Alina Hang, Daniel Buschek, Alexander De Luca, Axel Hösl,
Sebastian Löhmann, Henri Palleis, Simon Stusak, Sarah Tausch,
Emanuel von Zezschwitz, Andreas Butz, Heinrich Hussmann (Editors)

Special Aspects of Usability

Hauptseminar Medieninformatik SS 2014

Technical Report
LMU-MI-2014-2, September, 2014
ISSN 1862-5207

University of Munich
Department of Computer Science
Media Informatics Group

Alina Hang, Daniel Buschek, Alexander De Luca, Axel Hösl,
Sebastian Löhmann, Henri Palleis, Simon Stusak, Sarah Tausch,
Emanuel von Zezschwitz, Andreas Butz and Heinrich Hussmann
(Editors)

Special Aspects of Usability

Hauptseminar Medieninformatik SS 2014

Preface

This report provides an overview of current research dealing with special aspects of usability and covers various domains that deal with this topic, ranging from user experience, physical visualizations to usable security and privacy.

During the summer term 2014, students from the Computer Science Department at the Ludwig-Maximilians-University in Munich did research on specific topics related to 'special aspects of usability' and analyzed various publications. This report comprises a selection of papers that resulted from the seminar.

Each chapter presents a survey of current trends, developments, and research with regard to a specific topic. Although the students' background is computer science, their work includes interdisciplinary viewpoints such as theories, methods, and findings from interaction design, ergonomics, hardware design and many more. Therefore, the report is targeted at anyone who is interested in the various facets of 'special aspects of usability'.

Munich, September 2014

The Editors
Alina Hang, Daniel Buschek, Alexander De Luca, Axel Hösl,
Sebastian Löhmann, Henri Palleis, Simon Stusak, Sarah Tausch,
Emanuel von Zezschwitz, Andreas Butz and Heinrich Hussmann

Contents

Context-Aware Mobile Interfaces

Felix Manke

Abstract— Context awareness can help to improve usability in many areas of computing. Especially with modern, sensor-equipped mobile devices such as smartphones or tablets, a whole range of possibilities in context-aware computing opens up. In this work an extended definition of context is provided, which is tailored to these new devices and comprises location, time, individuality, activity, relations and environment. Additionally, sources of context are identified and different ways for accessing contextual data are presented. By examining a set of exemplary applications, different categories of context-based adaptations to mobile interfaces are determined. Finally, a critical discussion of mobile context usage, considering limitations and user privacy, concludes this work.

Index Terms—context, context-aware, mobile, mobile interface, adaptation

1 Introduction

With modern mobile devices such as smartphones and tablets becoming tremendously popular these days it has become completely natural for many people to carry around a computing device with them throughout the day. As a result of this many tasks such as reading email or surfing the web that have formerly been bound to stationary desktop computers are now carried out by users from any place at almost any situation. These situations and usage scenarios then each define a specific *context of use*.

Applications traditionally are designed to fulfil certain needs in a certain context. The interface and interaction are tailored to a specific group of users that will use the application under certain circumstances. However, with a mobile device these circumstances - the *context of use* - can only vaguely be foreseen (if at all). As a researcher or developer one does simply not now, whether the user will run the application at home, in the office or while standing in the subway hanging on to the handhold with one hand, trying to operate the smartphone with the other. Preparing an application for all of these possible situations is a challenging task. If there is no perfect solution that fits all possible contexts, the application somehow needs to determine in which situation it is being used: it needs to become *context-aware*.

Modern smartphones are equipped with a variety of sensors that allow them to gather contextual information such as location, direction, movement and acceleration, light, sound, and such. Therefore there is a chance to take into consideration all of this data within an application in order to determine the current situation and use case. In addition, today's mobile devices are equipped with enough processing power to do extensive calculation on these data. On top of that, modern smartphones are connected to the world wide web and all its news, 3rd party services, social networks, and communities which can also serve as a source of contextual information.

This paper gives an overview of how mobile interfaces can access contextual data and how they can adapt to the current context to offer improved usability to the user. In order to set the ground for this analysis a definition of context is provided first that fits the field of mobile devices such as smartphones. Following that, as set of mobile applications that make use of contextual information is presented and reviewed in order to provide insights into how contextual data is gathered and used today. Afterwards there follows a comparison of the examined applications and an outlook on future opportunities and interesting fields of research with respect to context. Finally, there will be a critical discussion on chances and risks involved in context awareness.

- *Felix Manke is studying Human-Computer-Interaction at the University of Munich, Germany, E-mail: felix.manke@campus.lmu.de*
- *This research paper was written for the Advanced Seminar Mediainformatics, 2014*

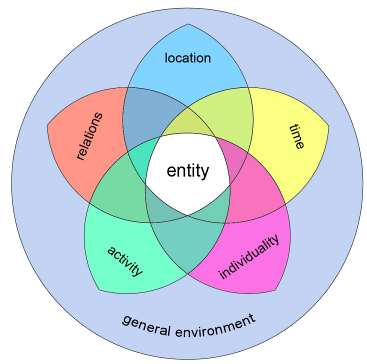


Fig. 1: The six aspects of context: (1) location, (2) time, (3) individuality, (4) activity, (5) relations, and (6) surrounding general environmental conditions. This definition extends the definition of Zimmermann et al. [28] by adding the encompassing dimension 'general environment' that represents the overall, natural environment (e.g. weather) and allows to further describe the context of an entity.

2 Definition of Context

Context and *context awareness* in Computer Science were introduced first by Schilit and Theimer in 1994 [23]. Back then the authors declared context to comprise the elements location, people, and objects. In the same year Schilit et al. identified three key aspects of context, defined by *'where you are, who you are with, and what resources are nearby'* [22].

With respect to *'ultra-mobile computing'*, characterized by devices that are operated on the move (e.g. mobile phones), this definition was developed further by Schmidt et al. to also include activity and the inner state of the user [24]. In 2001 Dey published his often cited definition of context as *"any information that can be used to characterize the situation of an entity"* [2].

Six years later Zimmermann et al. provided an operational definition of context [28]. The authors complained about existing definitions either being not precise enough or not generalizable, because most of them simply defined context by example or by synonym. By the definition of Zimmerman et al. the context of an entity consists of five elements: *location*, *time*, *individuality*, *activity* and *relations*.

Taking into account further, more recent work in the field of mobile computing context (such as the notion of 'virtual sensors' by Lovett and O'Neill [10] or the dependency of app usage on places and social context discovered by Do et al. [12]) this work extends the definition by adding a sixth, encompassing element: *environment*.

With today's highly connected smartphones (e.g. via the web) this sixth aspect of context can for example be accessed via 3rd party services that provide access to databases with (near) real-time data from all over the world (e.g. weather agencies). The six aspects then define the context of a given entity (user/device/object) for a certain moment *(see Figure 1)*. Together, they fully describe the entity's situation, while each of the six aspects captures some partial information:

1. **Location:** information such as position, speed and orientation
2. **Time:** the current time, time intervals, past and future events
3. **Individuality:** the user's preferences, interests, capabilities and behavior
4. **Activity:** current tasks, goals and needs that are expressed by actions
5. **Relations:** relations to other persons, objects, services and information
6. **Environment:** information about the natural, environmental context, such as weather, smog levels in cities or pollination that may affect the user but cannot directly be influenced or interacted with

With this definition of context in mind, what does it mean for an application to be *aware* of context? **Context-Awareness** basically is the ability to gather any relevant information that can characterize the environment of an entity and also deliver this information timely to applications [1]. For the application itself this means to gather any relevant contextual information about the six aspects of context (individuality, location, time, activity, relations, environment) that suffices to fully characterize the situation to an extend necessary to provide the intended service. Thus it depends on the application, what contextual information is needed and what date can be omitted. In reality however, it is usually not possible to really gather *all* desired information about the context of use. Therefore most of the time an application will need to deal with only a partial description of the relevant context.

3 Context-Awareness in Mobile Interfaces

Contextual information can be useful in a quantity of situations as it can support an application with data about the context of execution beyond the traditional information provided by the operating system.
Thereby, context can help in making decisions based on the current situation [11, 16], mitigate situational impairments [6] and augment other data with additional information - be it in common mass-market applications such as modern smartphones creating meta-data for photos (containing information like location and time), or be it in highly specialized fields such as health care [17, 25].

While there are many application areas for the use of contextual data, this work focuses on the possibilities for mobile interfaces to work with and adapt to the given context.
Adapting user interfaces according to the context of use aims at improving usability and enhancing user experience. This is achieved mainly by optimizing the interaction and reducing the error-rate and time needed for fulfilling certain tasks.

Yet, there are many different ways to adapt an mobile application interface to the given context to achieve the aforementioned goals. Therefore, a variety of applications will be examined in this section to get insights into the use of context in this field. After this there will be an analysis of potential sources of contextual information based on the examined applications followed by an overview of possibilities of how to access this information.

3.1 Categories of Context-Aware Mobile Interfaces

In this section a set of applications using context will be examined. They are grouped by categories which are defined by the way the contextual information is utilized. Some applications augment existing data with additional contextual information while others display the current context directly to the user. Moreover, there is the possibility to provide a context-based selection of data to the user, adapting the displayed information or the user interface visually, and also adapting the user controls to the context. Finally it is possible to change the status of the mobile device itself based on the user's situation.

3.1.1 Contextual Augmentation

One common application of contextual data is the augmentation of other data at the time of creation. For example it is a *de facto* standard nowadays, to add meta-data containing time and location to photos made with the camera in smartphones but also with stand-alone digital cameras equipped with GPS sensors[1]. This information can then for example be used to display and group the augmented data according to their embedded contextual information *(see Figure 2)*.

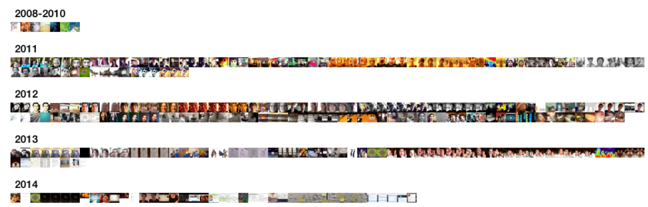


Fig. 2: Grouped display of photos with in the iOS 7 photo gallery , based on creation year.[2] This allows the user to quickly access pictures based on their historical time context.

This can be regarded as *retrospective* context-awareness as the contextual information does not resemble the current but past situations. The thereby created context history on the other hand can be useful in comparing the current context with past ones and can thereby help to identify specific situations.

Displaying additional, contextual information in *real-time* on the other hand has become popular and familiar by now due to all kinds of navigation systems displaying the current position and other metrics like speed or direction to the user.

Preuveneers and Berbers describe the use of mobile phones for personalized health care assistance [17]. By monitoring the location and activity of persons diagnosed with diabetes they can augment data logs of blood glucose levels with contextual data and are able to recognize certain behavioural patterns. As a result of that the application can assist the patient with taking well-informed decisions on drug dosage based on their current situation and activity. The position of the user is determined based on the acquisition of GSM cellular data. This provides only vague location data but is sufficient for recognizing locations that the user previously visited. Combined with a timestamp, logs of food intake, activities, measured glucose level and insulin dosage the location data define a specific situation that gets stored to a history for later comparison.

For the calculation of proper insulin levels the authors then try to find in the data history a situation (i.e. data set) which most resembles the current context. A selection of matching entries gets filtered and sorted and then presented to the user *(see Figure 3)*.

Fig. 3: Measuring the blood glucose level and providing a visual representation on a mobile phone for comparison with previous results. *(figure from: [17])*

[1] Examples of GPS-enabled digital cameras for automated location-tagging: Nikon Coolpix P330, Canon PowerShot SX280 HS, Panasonic Lumix DMC-TZ61, Sony Cyber-shot DSC-HX400V

[2] see: http://www.apple.com/ios/features/#photos

Another way of augmenting the mobile interface through contextual information is presented by Oulasvirta et al. with ContextContacts [16]. As a re-design of the smartphone's contact book, ContextContacts enriches the interface with contextual information about the current situation of the user's contacts. This is achieved by providing automatically generated, meaningful icons next to each contact in the list, which display information about the current situation of the corresponding person *(see Figure 4)*.

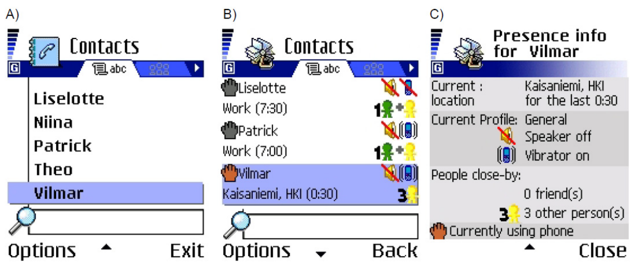


Fig. 4: ContextContacts: the standard contact book (A), the addition of contextual information about each contact (B), detailed information about a selected contact (C). The additional contextual information allows the user to decide, whether it might be appropriate to call someone or not. *(figure from: [16])*

This additional information can help the user in deciding, whether calling another person may be appropriate at that moment or not. The additional visual cues represent context information such as the location of the other person as well as the time he/she spent at this location. Additionally, the alarm profile of other users' phones can be viewed to see, whether they have muted their phone (which would indicate that they do not want to be called if not inevitable).

Another interesting cue is the indication of responsiveness in terms of phone usage by other users. A hand icon is displayed next to each contact which is turning from gray to red if the phone has or is being used. This way the user can see, whether a person in their contact list is currently using their phone actively or whether it may rest somewhere on a shelf unnoticed.

Finally, community-related information is provided as well. ContextContacts displays an indication of nearby people next to each contact-list entry. This takes into account whether the user knows the persons that are with the corresponding contact or not. Other persons known to the user are indicated by a yellow person icon while persons unknown to the user are indicated by a green person icon. The number of persons is displayed textually next to each icon.

Altogether the augmentation of the contact book with contextual information shall support the user by allowing him or her to estimate the current situation of others.

Lindqvist and Hong present a prototype application for Android devices to reduce this distraction caused by mobile phones while driving [9]. Distracted driving can lead to unnecessary accidents and human casualties. While there are many potential distractors in a car and on the road, one of them are mobile phones. When using their mobile phone while driving, a person can not fully concentrate on the road any more.

The system that Lindqvist and Hong present is context-aware and also implements the concepts of burden-shifting, time-shifting and activity-based sharing. Their application helps to reduce the need to operate the mobile phone while in the car based on five principles: (1) give callers information about the callee's context automatically, (2) make callers wait for the right time to make the call, (3) respond to incoming calls automatically where possible, (4) defer messages and send pre-planned messages, (5) share the estimated arrival time to other persons when on the road.

To achieve this, the application provides simple cues about the context of other users: available, busy and unknown. Additionally it provides hints at the reason for the status (e.g. 'driving') *(see Figure 5)*. This information is pushed to the caller, who is currently not driving and can take the cognitive load of making the decision about whether it is really necessary to make the call or not (burden-shifting). If the caller decides to contact the call recipient although he or she is driving at that moment, there are different options to choose from. The caller can for example decide to send a text message. If the message recipient is still driving, the message delivery will be delayed until the driver has reached his/her goal (time shifting). Another possibility is to call the voice-box or to set up a reminder for a later call. There still is the option to actually call the driver in case of an emergency. Further, a user can share his/her current status with another person for arranging later calls at an appropriate time. This sharing of information can also be driven by contextual factors. For example the user can automatically share his location to a certain person whenever he is driving (activity-based sharing).

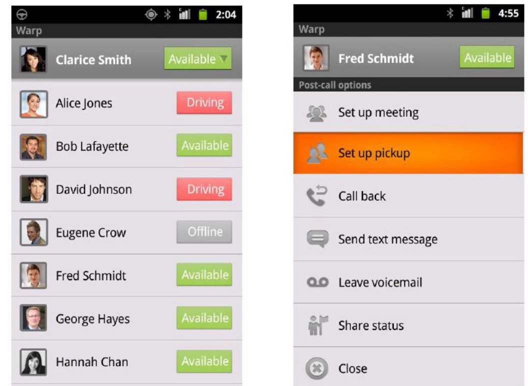


Fig. 5: Undistracted Driving: context-based information (left) and appropriate actions proposed based on the current situation of other users (right) *(figure from: [9])*

The described augmentation of data with contextual information allows for dynamic grouping, automatic tagging and later inspection and insights. Further, it can help in making decisions based on the current context. This can also hold true for the context of other users which can be made visible to the user. Altogether, the augmentation of data with context can be helpful in the organization and filtering of data as well as in making decisions based on the given context. This usage of context can be subtle, if the user does not directly notice or see that context is being taken into account (as with grouping and filtering) or more visible if the contextual augmentation is visualized directly as textual or pictorial information.

3.1.2 Context-based Options

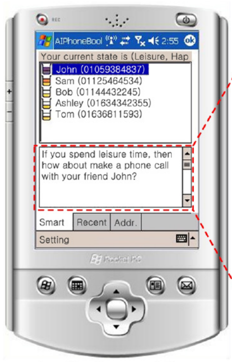
Another, less subtle use of contextual data is to directly affect the options presented within the application's user interface. This can be used to either provide the user with extra options or to hide information and choices which are unnecessary at that particular moment. Google's *'Local'* service[3] for example allows for searching interesting places nearby. It takes into account the user's current location to provide search results that are in spatial proximity to the user.

Hapori [8] is a local search technology for mobile phones that takes into account even richer contextual information. Local search engines are used to find nearby restaurants, shops or cafes on the go, typically using a GPS-sensor-equipped device such as a modern smartphone. The vision of the creators of Hapori is that future search queries should include much more information than just the query string and the current location. Therefore rich contextual information about the day of the week, time of day, current weather and weather forecast, and even the current activity of the user are included. Further the community of users with similar interests and behaviour is taken into account. Hapori therefore creates behavioural models of users and looks for similarities with other users in the community. This information then is

[3] see http://www.google.com/+/learnmore/local/

used to provide better search results compared to static, location based points of interest.

Min et al. have developed an application that recommends phone numbers of contacts based on the user's current situation [11]. It provides an ordered list of contacts, sorted according to a matching-score which is visualized as colored icon next to each contact's name. Additionally, textual information is provided when the user selects one of the proposed contacts *(see Figure 6)*.



Application prototype

Category	Script	Prior contexts
Anniversary	Today is Ashley's birthday. Why don't you make a phone call to her? ...	Day, schedule, related person, amity
Schedule	You have a meeting today 2 p.m. with John. You'd better make a phone call to him. …	Time, day, schedule, location, related person
Location	Your friend Bob is close to here. Do you want to make a phone call to him? …	User's location, local information, amity
Statistical pattern	You have not contacted to Ashley recently. Why don't you make a phone call to her? …	Recent call, recent SMS, amity
Etc.	Here is the phone number of the nearest car center. …	Device condition, news, …

Recommend scenario

Fig. 6: Context-aware contact recommendation *(figure from: [11])*

Min et al. derive the user's context from mobile logs using Bayesian networks (BNs). The log-data they analyse comes from different sources such as the personal information management system (PIMS) and the global positioning system (GPS). Also the user's calling history and short messages services (SMS) are taken into account. This allows for the extraction of three kinds of high-level contexts: (1) a social context, which describes the amity between the user and other persons in the social network, (2) the user's emotion and (3) the degree of busyness. Additionally, four kinds of low-level contexts (time, day of week, location, and schedule) are considered to model the user's service consuming behaviour. Proper services (here: calling specific contacts of the user's phone) can then be provided to the user based on the semantic compatibility between current and past contexts.

Another application providing services based on the user's context is CRUSE by Hodjat et al. [7]. CRUSE is a context-aware user interface framework for delivering applications and services to the user in a standard usable form. It allows to navigate between different applications and provides a consistent look and feel across applications integrating the framework within their user interface.

The framework aims at maximizing the amount of relevant information and minimizing the difficulty of accessing further information not displayed at that moment. To achieve this, CRUSE presents options (applications and services) that are most likely to be useful to the user in the given situation. This likeliness is derived from the available context which also takes into account user preferences and the user's behaviour (activity). As the prediction made based on the different sources of information can not be foolproof, the authors also provide a search box for natural language input. Here the user can type in queries and intended actions in its own words. The basic idea of the authors is that if an desired option is not available there has to be an easy way to ask for it: *'If you see it, click it. If you dont, ask for it.'* [7]. CRUSE also assists the user in the process of searching for further options by providing useful and meaningful suggestions and making proposals for search terms based on the user's searching history.

Xu et al. have worked on a context-aware app usage prediction model [27]. In order to improve device usability by for instance preloading applications that are most likely to be used next, they created an app prediction framework that works based on three key aspects: (1) the user's preferences and app usage history, (2) sensor based information about the environment, (3) community based pattern aggregation. With their framework they are able to provide personalized prediction that also relies on community behaviour to further tailor the result. The community of like-minded users thus becomes part of the current context: if the user is part of a community that makes heavy use of social media, this influences the applications he/she will use on a regular basis. Exploiting the community ties can result in more robust predictions as the community-context can compensate in situations where personal context-information may be thin.

Altogether, the presentation of options based on the current context can help to provide user interfaces with better usability by creating a less cluttered experience. This is mainly achieved by leaving out unnecessary elements while visually focusing on the important ones or by filtering options based on the user's current situation. This can be done on a per-application-level but also on the device-level to provide fitting applications or services at the right place and at the right time. The overall goal of this type of context-driven adaptation is to assist the users with the interaction and to take some of the selection work off their shoulders.

3.1.3 Visual Adaptation of Displayed Information

Besides the augmentation of data and context-based selection of information, the displayed information can also be visually adapted to the user's current context. A common example for this is the 'night mode' of navigation systems that gets activated automatically when the sun sets *(see Figure 7)*.

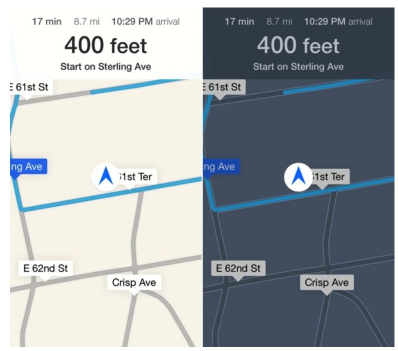


Fig. 7: Maps application for Apple iOS 7: daytime navigation (left) and night mode (right) [4]. Contextual information (time) is taken into account to adapt the interface in a way that is beneficial to the user at that particular moment.

There are more examples for context-dependent adaptation of visuals and researchers are looking for new ways to adapt the displayed information visually in context-aware applications.

With CAMB Gasimov et al. present a context-aware mobile web browser [4]. The goal of their work was to create a framework for effective context-aware mobile web browsing. As today's mobile phones provide the user with the possibility to surf the web and are taken to different places as well as used in many different situations, the authors identified the need to create a system to adapt web pages to the context of usage.

According to Gasimov et al., for web browsers on mobile devices there are two types of context: static and dynamic. Static context does not change during the session (for example the operating system, phone manufacturer, web browser, screen resolution). Dynamic context on the other hand can change during the session (for example battery, location). Therefore, the concept of CAMB allows for the adaptation of the layout and presentation of a websites' elements according to the (static) device properties and also the user's environment, his/her mood, and current activity. The sources of context are the static and dynamic device information (operating system, battery level, sensor data) but also information coming from 3rd party providers such as the current weather forecast.

The change in a websites' appearance is realized by using the web's standard style sheet language CSS (Cascading Style Sheets). By mak-

[4] see: http://www.apple.com/ios/maps/

ing use of the 'media'-property of CSS the authors can provide special styles for situations like 'walking' or 'running'. Originally, the 'media'-property has been designed to provide individual styles according to properties of the presentation medium like the screen's size or for special use cases such as printing. But, according to the authors, there is 'no technical barrier' to use 'media' for the definition of context-dependent styles as well.

CAMB implements three ways to adapt a website's content (i.e. text, images, video, and animations): (1) changing the placement and order of elements, (2) showing or hiding certain elements based on the context, (3) adjusting display properties of elements (e.g. the size of the font). An example of this adaptation can be seen in Figure 8.

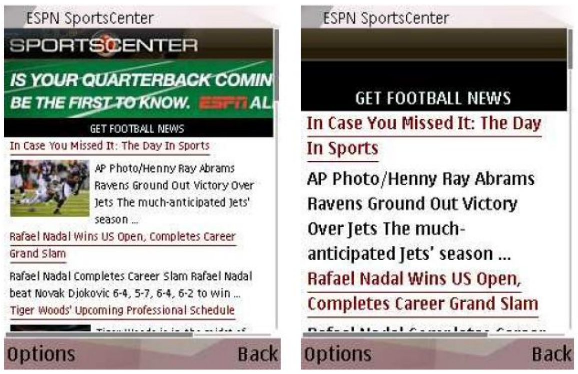


Fig. 8: Example of the adaptation of a web page made by CAMB: original page (left), adapted page for walking context (right) *(figure from: [4])*

An earlier approach to adapting web documents to the user's context is presented by Nathanail et al. [13]. Their system uses contextual information about the user environment, gathered through a wireless sensor network, to adapt the presented content to the current environmental conditions.

Altogether, systems that adapt the visualization of the user interface's elements to the current context mainly help to improve perception and readability of the displayed information. This can be useful for different environmental conditions (bright/dark, loud/quiet, crowded/empty) as well as for different user activities (user at rest, walking, running). While most of today's applications reduce the adaptation to more conservative changes like different color themes, also more radical changes, as completely re-ordering content can be imagined and have also been examined by researchers.

3.1.4 Adaptation of User Controls

Aside from data and visual elements also the user controls can be adapted. As a modern smartphone only consists of a single large touch screen, most of the controls also underlie the influence of the application as they are only virtual, being rendered to the device's screen (e.g. the virtual keyboard).

ContextType by Goel et al. is an adaptive system for text entry that leverages hand posture information to improve touch input on mobile devices [6]. Mobile text entry is error-prone due to the fact that a mobile phone gets carried around with the user and is used in many different situations at different places throughout the day. This can lead to different ways of text input (e.g. with two thumbs or with one digit only). The authors render the example of a user standing on a moving bus, supporting himself with a grab-handle with the dominant hand such that only one hand - the non-dominant one - is left for holding the phone and typing at the same time. The system switches between different virtual keyboard models to dynamically adapt to the way the user inputs text. This adaptation does not affect the visual layout of the keyboard though. Instead the underlying touch recognition model changes to improve the rate of correctly recognized keys. The touch input information is then combined with a language model to make the entry of text more accurate.

For that, ContextType infers three types of information: (1) the user's current hand posture, (2) the user-specific touch pattern (gathered by a training application), and (3) letter probabilities from a language model. The system then modifies the motor-space location of each key according to the user-specific touch input behaviour *(see Figure 9)*.

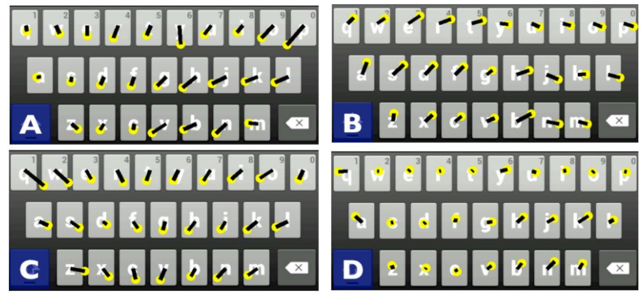


Fig. 9: Example of different touch patterns for ContextType: (A) Left Thumb, (B) Right Thumb, (C) Index Finger, (D) Two Thumbs. The yellow spot indicates the touch centroid for each key, the line visualizes the drift from the visual key center. *(figure from: [6])*

In a first evaluation the authors found that ContextType reduced the total error rate by 20.6%.

The context-based adaptation of user controls is particularly interesting as there are the same issues as with visual presentation: depending on the situation a mobile phone user may have to rely on completely different input methods. Besides the rather subtle adaptation of the keyboard model by Goel et al. [6] also more radical changes could be thought of (e.g. switching to voice input or controlling the device by gestures). Overall the adaptation of user controls aims at overcoming situational impairments and to improve the usability by reducing the input error rate.

3.2 Accessing Contextual Data

As there are many different application areas for context-aware mobile applications (examples are examined in the previous section) that differ strongly in terms of what they provide the user with and how they achieve this, it can be valuable to take a look at the different sources of contextual data. This section therefore gives an overview of the common sources of context and also takes a look at different possibilities for applications (and consequently developers and researchers) to access these valuable contextual-data.

3.2.1 Sources of Context-Data

As context is a complex phenomenon that consists of more or less aspects (depending on the definition and the application area) also the possible sources of context are diverse. However, as there are no standards established yet in application development environments, platform architectures and services across different mobile devices, it is (as of the time of writing) usually a difficult process to develop context-aware applications [3].

As set out above, this work defines context as the combination of five key aspects: (1) **individuality** (e.g. the user's preferences and behaviour), (2) **time** (e.g. the time of the day, the day of the week) (3) **location** (i.e. at home, at work, at the Eiffel Tower), (4) **activity** (e.g. jogging, driving, sleeping), (5) **relations** (e.g social communities, nearby persons, devices, services) (6) **general environment** (e.g. weather, smog, tide).

Nevertheless, these aspects of information are not directly visible to an application. Any application - be it on mobile devices or on stationary systems - needs to somehow gather data that allow for the deduction of this information. From the examples given in the previous chapter the following possibilities can be extracted:

1. **Location:** GPS, mobile and wireless network data, manual user input ('home', 'office')

2. **Time:** device clock, calendar
3. **Individuality:** user-specific device settings and preferences, user profiles/accounts, activity and application usage-logs, status messages
4. **Activity:** manual user logs, sensor data, phone state, alarm mode
5. **Relations:** other devices within wireless network (e.g. Bluetooth, WiFi), sensor data (e.g. audio), network connection (e.g. web services), social networks and communities, manual user input
6. **General environment**: 3rd party providers (e.g. weather agency), manual user input

Another question is how to extract useful information from contextual data. The location by GPS in longitude and latitude for example is not useful at all if there is no point of reference provided. This point of reference could be a digital map or some other places given in longitude and latitude (for example locations previously visited by the user). In addition the location context can be defined as the *absolute position* or as the *relative position to some other location*. Further it is possible look at the physical location in a quantitative/geometric way (e.g. numerical distance) or in a qualitative/symbolic way (e.g. at home, at Anna's place).

As a consequence, depending on what is important, different aspects of the given data need to be considered. Most applications work out what they need on their own - gathering raw data from sensors - based on the specific usage scenario. This can be a tedious process and also means, that the same work has to be done again and again by different developers and researchers. In most cases, the goal of every application is to identify *situations* from contextual data and adapt the interface accordingly. However, what humans can describe easily (e.g. 'sitting in the subway, very hungry, five minutes behind schedule, hastily reading my last emails and nervously worrying about the growling, smelly dog of the elderly man sitting opposite to me') is hardly to perceive for an electronic device. Nevertheless, exactly this kind of highly descriptive information would be most valuable to an application. Until now, an application can try to determine whether a fast moving user is jogging or riding a bike by analysing the data coming from the accelerometer: constant vertical up-and-down movement accounts for jogging, rather smooth movement and more or less no vertical movement could indicate a bike ride. It is difficult to maintain such software and adding new context information usually requires a modification of the whole application which also means redeploying it to the end user. So, instead of every application trying to determine this kind of information on their own, general-purpose context providers would be desirable.

3.2.2 Context Providers

Instead of directly accessing the contextual raw data coming from sensors and other sources of information, there is another possibility: access to contextual information via special middleware or frameworks that mediate context between the hardware and the application and provide abstract, transparent access to context data.

This way an application can concentrate on situations - which is the essence of context data - that the user is in. In general an application needs contextual data to derive the user's situation. It is important whether he/she is working, sleeping, driving in the car or jogging in the park. Moreover it may be important to know, whether there are other people around and what relationship they have with each other. Most of these facts can be described by an abstract situation: driving, in a meeting, in the subway, at home with the family on a rainy evening and so on. There are approaches to provide exactly this kind of abstraction and this section introduces a selection of them.

An early approach on general context provisioning has been presented by Salber et al. in 1999 [21]. As part of the Context Toolkit, Context Widgets, a widget library for sensing context information (presence, identity and activity of people/things) has been developed. These context widgets mediate between the environment (i.e. the data coming from different sensors and other sources of contextual information) and the application. The idea is based on the concept of graphical widgets that provide an interface for the user to interact with an application. As toolkits and widget libraries have been very successful with the creation of graphical user interfaces, the authors suggest that widgets for accessing context and integrating it into an application could help building context-enabled applications faster.

Gellersen et al. described in their work [5] an approach to augment mobile phones with context-awareness by attaching them to a self-contained module (plug-in). The TEA (Technology Enabling Awareness) module measures the user's context with several sensors simultaneously and automates profile activation (e.g. for the situations in-hand, on-table, in-pocket, or outdoors...). The authors wanted to provide an alternative to the use of single powerful location- and vision-based sensors and instead combine many comparatively simple sensors to create awareness of situational context. They were convinced that multiple sensors that each capture only a small aspect of the environment could - when combined properly - render a total picture of the device's surrounding that *'better characterizes a situation than location- or vision-based context'* [5].

Another platform that is developed as context provider on mobile devices is ContextPhone [18]. It can sense, process, store, and transfer context data. Additionally, it integrates and interfaces with existing solutions on the smartphone such as the messaging and calling functions. Thus, actions can be triggered based on contextual cues or the inferred context can be communicated to the other users or applications.

Vihavainen et al. introduce ContextLogger3, a context logging-tool that combines data collected from sensors and phone activity with user created textual notes [26]. As potential target audience the authors identified end-users, service providers, 3rd party services, and researchers. The motivation for the user-generated textual notes is, that the automatically sensed context is only of limited precision. As mentioned before, sensors currently only deliver a very incomplete representation of the surrounding world. Further, the user's emotions and personal impressions can not clearly be captured technically. To 'close the gap' between the reality sensed by the sensors and the one being perceived by the user, the textual notes have been added to the logging tool.

MoReCon (short for Mobile Restful Context-Aware Middleware) is a web service that makes context information accessible to all kinds of applications by providing a RESTful API (Application Programming Interface) [3]. MoReCon thereby creates a layer between the aggregation and the management of context information from the mobile device's sensors and other sources of context and the application itself. It acts as a middleware to abstract this management of context data from the context-aware application and to provide a standardized way for accessing contextual information on all types of platforms. MoReCon provides an universal HTTP API for accessing context data using web services.

Ntawanga et al. present an integrated logical context sensor that determines context information in realtime to be used in various mobile web adaptations [14]. It was developed as a 'plug-in' component to be incorporated into mobile web applications for providing context awareness.

Other frameworks that deliver context awareness as a service to other applications are Zonezz [20], a platform for identifying meaningful locations (e.g. 'home' or 'work' instead of raw coordinates) and Controy [19], a middleware for efficient context provisioning on mobile devices.

It shows that there is much research going on in the field of context-provisioning. If context becomes a standard resource that can be accessed easily by all applications via simple and efficient APIs, also the adaptation of interfaces to specific situations will benefit from that.

4 Summary and Comparison

This section provides a summary and overview of all applications that have been mentioned and of the way they access context and adapt to it.

Application	Individuality	Time	Location	Activity	Relations	Environment	Type of adaptation
CAMB [4]	-	-	x	x	-	x	visual adaptation
ContextContacts [16]	x	-	x	x	x	-	context-based options
ContextType [6]	x	-	-	x	-	-	adaptation of user controls
CRUSE [7]	x	(x)	x	x	-	-	context-based options
Diabetes Health Assistant [17]	x	x	x	(x)	-	-	contextual augmentation
Hapori [8]	(x)	x	x	x	x	x	context-based options
Preference & Communities [27]	x	x	x	(x)	x	x	context-based options
TEA [5]	-	-	(x)	x	(x)	(x)	*depends on application*
Undistracted Driving [9]	-	x	x	x	x	-	context-based options

Table 1: The different aspects of context and their consideration by the examined applications. The table shows that the applications vary greatly in terms of what sources of context are taken into account and what types of adaptations result from the additional contextual information. It can also be seen, that only a few applications make use of the full range of the six aspects of context.

There have been identified four areas for using context in mobile applications and interfaces: (1) contextual augmentation, (2) context-based options, (3) visual adaptation of the displayed information, and (4) the adaptation of user controls.

On a meta-level, context data can also be used to adapt the device's functionality or state and thereby the interface as a whole. For instance it is possible to change the user profile or the operational mode (e.g. flight mode) [5]. However, the changes made here, generally result in contextual adaptations that fall into one of the four categories mentioned above.

Table 1 compares all applications examined in this work with respect to the sources of context they utilize. The table shows that (a) the applications strongly vary in their set of utilized sources of context, that (b) the most commonly used contextual information is activity and location, and that (c) only two of the applications take all aspects of context into consideration. Further, the adaptations made based on the additional contextual information are different from application to application.

While this work does neither pretend nor attempt to give an exhaustive and complete overview of all existing applications, the comparison of the given example applications illustrates how different the current approaches to context awareness are - both in terms of sources of contextual data and the type of adaptation.

One of the most common use cases for context nowadays is the selection or presentation of context-based options to the user. Also, the augmentation of data with contextual information is a popular way to create additional meta-data (e.g. for photos). A very interesting but rare application area is the adaptation of user controls. This is particularly attractive, as different situations require different input methods and as every user displays slightly different input patterns and behaviour.

5 Discussion

Where does context-awareness make sense? As modern mobile devices and especially modern smartphones are equipped with a wide range of possibilities for gathering contextual data it may be tempting to make every application context-aware. However - as seen from the examples - not every context-aware application makes use of the full range of possibilities. Furthermore, as will be highlighted by this discussion, context-awareness is not always useful. This is due to the fact that (a) the detection and recognition of contextual data is not trivial when it comes to more complex correlations and (b) not every information is useful for all applications. Further, the constant sensing of contextual data needs additional energy which is generally limited with mobile devices. Finally, the extraction of meaningful cues from the contextual raw data can be computationally intense and also produce large amount of data if to be stored in a context history.

Nevertheless, for certain applications, context is a very valuable source of information that can be used to greatly improve usability. However, today's technical possibilities not always succeed in fulfilling the desires of researches and developers. One of the main problems is, that the collected contextual data can not fully represent the user's situation. While everything that is happening directly to the device can be sensed rather precisely, information about the environment usually is rather vague. It is simply not possible to capture a holistic picture of the surroundings (nearby devices, people, services, properties and characteristics of the physical environment, etc.).

An interesting field of research would be to increase the networked interoperability of different devices. This could for example allow to recognize (or estimate) how many people are around based on their smartphones communicating their presence to each other.

Besides, the user's mental state (emotions, intentions) can currently not be captured directly. Yet, this today missing information would open up completely new possibilities.

Additionally, it would be desirable to generalize context recognition. The recognition of all contextual information generally gets merged to identify a specific situation. As most applications try to identify situations based on the context to provide the user with exactly what he/she needs at that moment, it would be helpful to have a central instance for context recognition on the device. If for example the mobile device's operating system would identify situations and provide them directly to application developers (as with battery state, user preferences, etc.) it would be much easier to work with contextual information.

One thing that must not be forgotten when talking about contextual information is privacy. This becomes clear especially with applications that share the user's context with other devices or persons. An example of a possibly critical amount of shared information is given by ContextContacts [16]: the application displays not only the location of other users but also how much time they spent there so far. This can be something, the user does not always want others to be able to see. The more information about the user's current situation gets collected, the more personal and sensible details can be deducted. Applications using context have to find a balance between context and security. Who can access the user context? How can respective rights be managed? Where does the information get stored?

Context can definitely help to improve usability and comfort when interacting with mobile interfaces. However, context alone can not produce good value and more context does not automatically mean better applications. The data about the current situation, relations, the environment and activity should only be used when the application (and ultimately the user) really benefits from it. Otherwise unnecessary resources in development and while running the application may be wasted and the user may even be overwhelmed by contextual output from the device. Even worse, the simple display of contextual information alone may be ignored completely by the user (as observed by Oulasvirta et al. [15]). In general it can be said, that contextual adaptations work best when subtle and straightforward: *access context information only if the user can be expected to benefit from it in a substantial way.*

6 Conclusion

This paper has examined several context-aware mobile applications to identify the possibilities for mobile interfaces to adapt to the user's current context and situation. There have been identified six conceptual sources of contextual data (individuality, time, location, activity, relations, environment) and an extended definition of context, comprising all six aspects has been provided. There have been identified four general ways to adapt a mobile interface to context: (1) augmentation of existing data and interface elements with contextual information, (2) the provisioning of context-based options to the user, (3) the visual adaptation of the interface or certain interfaces elements to the current situation or activity, and (4) the adaptation of user controls.

Further, two ways to gather contextual information have been identified: (1) directly from within the application by evaluating and analysing data from sensors and other sources of context and (2) from a context-middleware that mediates between hardware and application and provides abstract, transparent access to context data (e.g. in form of abstract 'situations'). In the discussion, critical points regarding context have been presented and also interesting fields for future research have been identified.

References

[1] P. Bellavista, A. Corradi, M. Fanelli, and L. Foschini. A survey of context data distribution for mobile ubiquitous systems. *ACM Computing Surveys*, 44(4):1–45, Aug. 2012.

[2] A. K. Dey. Understanding and Using Context. *Personal and Ubiquitous Computing*, 5(1):4–7, Feb. 2001.

[3] E. Dogdu and O. Soyer. MoReCon: A Mobile Restful Context-Aware Middleware. In *Proceedings of the 51st ACM Southeast Conference*, ACMSE '13, page Article No. 37, Savannah, GA, USA, 2013. Copyright 2013 ACM.

[4] A. Gasimov, F. Magagna, and J. Sutanto. CAMB: Context-Aware Mobile Browser. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, MUM '10, page Article No. 22, Limassol, Cyprus, 2010. ACM New York, NY, USA 2010.

[5] H. W. Gellersen, A. Schmidt, and M. Beigl. Multi-Sensor Context-Awareness in Mobile Devices and Smart Artifacts. *Mobile Networks and Applications*, 7(5):341–351, 2002.

[6] M. Goel, A. Jansen, and T. Mandel. Contexttype: using hand posture information to improve mobile touch screen text entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2795–2798, Paris, France, 2013. Copyright 2013 ACM.

[7] B. Hodjat, S. Hodjat, N. Treadgold, and I.-M. Jonsson. CRUSE: A Context Reactive Natural Language Mobile Interface. In *Proceedings of the 2nd annual international workshop on Wireless internet*, WICON '06, page Article No. 20, Boston, MA, United States, 2006. ACM New York, NY, USA 2006.

[8] N. D. Lane, D. Lymberopoulos, F. Zhao, and A. T. Campbell. Hapori: Context-based Local Search for Mobile Phones using Community Behavioral Modeling and Similarity. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 109–118, Copenhagen, Denmark, 2010. ACM New York, NY, USA 2010.

[9] J. Lindqvist and J. Hong. Undistracted Driving: A Mobile Phone that Doesn't Distract. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, HotMobile '11, pages 70–75, Phoenix, Arizona, USA, 2011. ACM New York, NY, USA 2011.

[10] T. Lovett and E. O. Neill. Mobile Context-Awareness: Capabilities , Challenges and Applications. In *Adjunct Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing - Adjunct*, Ubicomp '10, pages 539–540, Copenhagen, Denmark, 2010. ACM New York, NY, USA 2010.

[11] J.-K. Min, H.-T. Kim, and S.-B. Cho. Social and Personal Context Modeling for Contact List Recommendation on Mobile Device. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '08, pages 381–384. IEEE Computer Society Washington, DC, USA 2008, Dec. 2008.

[12] T. Minh, T. Do, J. Blom, and D. Gatica-Perez. Smartphone Usage in the Wild: a Large-Scale Analysis of Applications and Context. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pages 353–360, Alicante, Spain, 2011. ACM New York, NY, USA 2011.

[13] S. Nathanail, V. Tsetsos, and S. Hadjiefthymiades. Sensor-driven adaptation of Web document presentation. In *Universal Access in Human-Computer Interaction. Applications and Services*, pages 406–415. Springer Berlin Heidelberg, Beijing, China, 2007.

[14] F. Ntawanga, P. O. Box, P. Elizabeth, and L. Barnard. An Integrated Logical Context Sensor for Mobile Web Applications. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, SAICSIT '13, pages 320–330, East London, South Africa, 2013. ACM New York, NY, USA 2013.

[15] A. Oulasvirta, R. Petit, M. Raento, and T. Sauli. Interpreting and Acting on Mobile Awareness Cues. *Human-Computer Interaction*, 22(1):97–135, 2007.

[16] A. Oulasvirta, M. Raento, and S. Tiitta. ContextContacts: Re-Designing SmartPhone's Contact Book to Support Mobile Awareness and Collaboration. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, MobileHCI '05, pages 167–174, Salzburg, Austria, 2005. ACM New York, NY, USA 2005.

[17] D. Preuveneers. Mobile Phones Assisting With Health Self-Care : a Diabetes Case Study. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices & services*, MobileHCI '08, pages 177–186, Amsterdam, the Netherlands, 2008. ACM New York, NY, USA 2008.

[18] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen. ContextPhone: A Prototyping Platform for Context-Aware Mobile Applications. *IEEE Pervasive Computing*, 4(2):51–59, Apr. 2005.

[19] O. Riva. Contory: A Middleware for the Provisioning of Context Information on Smart Phones. In *Proceedings of the ACM/IFIP/USENIX 2006 International Conference on Middleware*, Middleware '06, pages 219–239, Melbourne, Australia, 2006. Springer-Verlag New York, Inc.

[20] J. Roth. Context-aware apps with the Zonezz platform. In *Proceedings of the 3rd ACM SOSP Workshop on Networking, Systems, and Applications on Mobile Handhelds*, MobiHeld '11, pages 1–6, New York, New York, USA, 2011. ACM Press.

[21] D. Salber, A. K. Dey, and G. D. Abowd. The Context Toolkit: Aiding the Development of Context-Enabled. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '99, pages 434–441, Pittsburgh, Pennsylvania, USA, 1999. ACM New York, NY, USA 1999.

[22] B. N. Schilit, N. Adams, and R. Want. Context Aware Computing Applications. In *Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications*, WMCSA '94, pages 85–90, Santa Cruz, California, USA, 1994. IEEE Computer Society Washington, DC, USA 1994.

[23] B. N. Schilit and M. M. Theimer. Disseminating Active Map Information to Mobile Hosts. *Netwrk. Mag. of Global Internetwkg., IEEE*, 8(5):22—-32, 1994.

[24] A. Schmidt, M. Beigl, and H.-w. Gellersen. There is more to Context than Location. *Computers and Graphics*, 23:893–901, 1998.

[25] M. B. Skov and R. T. Hoegh. Supporting information access in a hospital ward by a context-aware mobile electronic patient record. *Personal and Ubiquitous Computing*, 10(4):205–214, Oct. 2005.

[26] S. Vihavainen, K. Karhu, and A. Botero. Connecting stakeholders through context logging. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices & services.*, MobileHCI '13, page 504, Munich, Germany, 2013. ACM New York, NY, USA 2013.

[27] Y. Xu, M. Lin, H. Lu, G. Cardone, N. D. Lane, Z. Chen, A. Campbell, and T. Choudhury. Preference , Context and Communities : A Multi-faceted Approach to Predicting Smartphone App Usage Patterns. In *Proceedings of the 2013 International Symposium on Wearable Computers*, ISWC '13, pages 69–76, Zurich, Switzerland, 2013. ACM New York, NY, USA 2013.

[28] A. Zimmermann, A. Lorenz, R. Oppermann, and S. Augustin. An Operational Definition of Context. In *Proceedings of the 6th international and interdisciplinary conference on Modeling and using context*, CONTEXT'07, pages 558–571. Springer-Verlag Berlin, Heidelberg 2007, 2007.

Measuring User Actions

Stefan Lau

Abstract— In this paper we give an overview over currently employed methods to quantitatively model the user in HCI. We outline the history of user models in HCI and highlight current developments in this area. Furthermore we show how researchers extract models from experimental data and how they are applied to interface optimization problems. Finally, we discuss the current state of research and compare existing models. Our review identified a lack of arbitrary models for three-dimensional movements and we show directions for future research to close this gap.

Index Terms—User Model, Perception Model, Human Computer Interaction, Information Theory, Fitts' Law, Steering Law, Throughput

1 Introduction

Designing beautiful and effective interfaces has always been a challenge in human computer interaction. While the beauty of the interface is largely up to the user's opinion, effective interfaces are often achieved using guidelines and best practices. As an alternative, or complimentary to these qualitative measures, there is a quantitative way to describe how effective a user interface can be operated. Models that try to match the movement, capabilities and potential of humans can be used to predict how fast and accurate a user can interact with the computer. Furthermore this allows us to show how much information can be transfered between human and computer.

The importance of the study of human movement was already recognized early in scientific literature. In 1899 Woodworth studied the accuracy of voluntary movement, observing it from a psychologists point of view [38]. With the introduction of information theory by Shannon et al. the exchange of information was modeled mathematically as a communication channel for the first time [35]. This lead to an uprise of quantification of perceptual and motor functions. 'The information capacity of the human motor system in controlling the amplitude of movement.' by Fitts [15] was initially published with the intention of quantifying human motor capabilities in one dimension, but developed to be one of the most important models of human computer interaction. Other researchers also developed models for the rate of information processing based on the same foundations, most notably Hick & Hyman [19] who established a model for decision taking. Interestingly the adoption of these laws has been very divergent. Fitts' Law dominates the field of human models in HCI. Other models, for example, by Meyer et al. [26], who described a movement as composed of two sub-movements, were less successful. Also, rather than using Fitts' Law to measure the information capacity, it was also used to predict movement times for pointing tasks.

$$MT = a + bID = a + b\log_2\left(1 + \frac{D}{W}\right) \quad (1)$$

A common notation of Fitts' Law can be seen in equation 1, where MT is movement time, D is the distance to the target and W is the width of the target. The parameters a and b can be determined by linear regression and relate to throughput. The logarithmic part is commonly refered to as the index of difficulty (ID), and is expressed in the unit Bits. A concept of human computer interaction as a communication channel emerged. Researchers also adapted the existing laws to more complex modalities. MacKenzie et al. were the first to adapt Fitts' Law to two dimensions [24] to make it usable for current input systems in HCI. Simultaneously to these predictive models for human performance, descriptive models were developed that tried to show how users interacted with a device. One of these models was the GOMS model [10], which split a task into goals, operators (actions), methods (sequences of operators) and selection rules (when does a user apply which method). These models allowed to estimate a tasks complexity and whether it is easy to execute. Other models, like the Keystroke-Level-Model [10] tried to span the gap between predictive and descriptive models. KLM did this by allowing only certain operators and assigning execution times to these operations. The sum of the execution times for a task can be determined easily. In this paper we will focus on predictive models, because they allow for quantified optimizations of interfaces.

2 Modeling the User

Predictive models for human interaction focus on a single aspect of the user that can be modeled, allowing the complexity of the human biology to be broken down into simple and easy to use formulas. This allows the researcher to focus on the characteristic of the user that he deems important. When a more complex model of the user is needed, these partial models can also be combined [13].

2.1 Movement Time

In current research the most common goal is to model how long a user takes to finish a task. In human computer interaction, these tasks usually incorporate moving or pointing a cursor.

2.1.1 In Two Dimensions

A lot of research is currently focused on the prediction of movement time between a starting point and a target rectangle. This is due to most of the current input systems being two dimensional, target-to-target systems. Examples for these kinds of systems are touchscreens, as implemented in current smartphones, or the WIMP metaphor in current desktop environments.

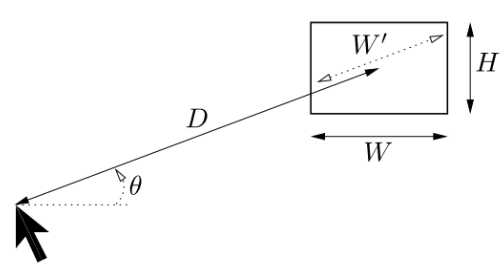


Fig. 1: Fitts' Law task in two dimensions. [4]

As stated in Section 1, MacKenzie et al. were the first to adapt Fitts' Law to two dimensions [24]. Figure 1 shows such a two-dimensional task. The participant started at the cursor and had to click the square of width W and height H as fast as possible. They compared several new definitions for the index of difficulty for two-dimensional tasks and extracted the best-fitting ones for their new model. To cater for the increased dimensionality, they introduced the apparent width W^I

- *Stefan Lau is studying Media Informatics at the University of Munich, Germany, E-mail: stefan.lau@campus.lmu.de*
- *This research paper was written for the Media Informatics Advanced Seminar 'Special Aspects of Usability', 2014*

which is calculated from W, H and the approach angle θ as seen in Figure 1. A minimal model using $W^{min} = min(W,H)$ was tested as well. Both of these models yielded approximately the same results for two-dimensional movements.

While MacKenzies model was very successful, only a few different configurations were used in their experiments. This was catered by Accot et al., who refined the models for two-dimensional pointing by assuming the task to consist of an amplitude and a directional pointing task [4]. Thus, they proposed a model that predicted the movement time for an amplitude pointing task when height tends to infinity and a directional pointing task when width tends to infinity. They proposed several euclidean norms that satisfy these constraints as new indexes of difficulty and tested their validity against their own data and data of other researchers. They concluded that the model best fitting all data was an euclidean model defining the index of difficulty as seen in equation 2. Here again D is the distance to the target and W and H the width and height of the target.

$$ID = \sqrt{\left(\frac{D}{W}\right)^2 + \eta\left(\frac{D}{H}\right)^2} \tag{2}$$

One disadvantage that Accot et al. introduced was the loss of the approach angle that was included in the W^I model. Therefore the approach angle did not influence the index of difficulty anymore. This issue was addressed by Grossman et al. [16]. They developed a probabilistic model based on the spread of hits when pointing at targets. This was accomplished by defining the probability of hitting a target in a two-dimensional task and calculating the indices of difficulty from these probabilities. The probability of hitting a target could also be calculated for non- rectangular targets, which allowed this model to be used for arbitrary shapes. For rectangular targets they assumed the hits to be normally distributed. A universal function F was derived from a one dimensional Fitts' Law ID, which maps probabilities to IDs. Equation 3 shows how to calculate an index of difficulty for a rectangular target using this universal function F and the probability of a hit of the target P_{Hit}.

$$ID = F(P_{Hit}) \tag{3}$$

While all of these models only regarded the distances in screen space, Chapuis et al. tested how the different scales affect Fitts' Law [11]. For this purpose they observed the influences of visual scale, motor scale and quantization (the fact that the cursor only moves in full pixels) on the models. They concluded that Fitts' Law did not hold up for targets that are small in motor or visual space and large in the respective other spaces. They proposed additional ID calculations for targets that are small in motor / visual space respectively.

Other approaches rather tried to adapt Fitts' Law to certain input modalities or test whether it was suitable for them. As touchscreen input became common with smartphones and tablets, Bi. et al. highlighted that Fitts' Law did not hold up correctly for finger input [7]. They found that Fitts' Laws predictive power decreased when targets got small. Additionally the finger touch is has an absolute imprecision. To account for that, they derived a new formulation for the index of difficulty based on a dual-distribution hypothesis. The distribution was built on a relative component, describing the speed-accuracy-tradeoff of the performer and a absolute component reflecting the absolute precision of the finger touch. The index of difficulty is calculated as seen in Equation 4 where σ and σ_a are the standard deviations of the dual-distribution and can be extracted from experimental data. D as always denotes the distance to the target. Unfortunately the two-dimensionality of the tasks was not respected in this model.

$$ID = log_2\left(\frac{D}{\sqrt{2\pi e\left(\sigma^2 - \sigma_a^2\right)}}\right) \tag{4}$$

Target-to-target movements are not the only type of movement that is common on current devices. Another type of movement is a trajectory-based movement, which allows movement in an corridor along a trajectory. An example for this kind of movements is a selection in a dropdown menu, or a gesture made by a finger. Figure 2 shows a trajectory based movement and the according corridor.

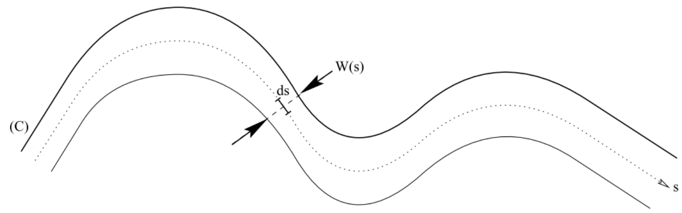


Fig. 2: Trajectory-based task in two dimensions. [1]

A model for these kinds of movements was first implemented by Accot and Zhai in 1997 [1] and is now commonly known as the Steering Law. The concept of the speed-accuracy tradeoff was continued in this model. They derived this model from a goal passing task, by increasing the number of goals and decreasing the distance between them. The basic equation was the same as Fitts' Law (still modeling the interaction as a speed-accuracy tradeoff) but an integral over the width of the tunnel was used as the index of difficulty formulation as seen in Equation 5. The findings were validated by multiple experiments that were executed by 13 participants. Additionally it was possible to model the speed of the user at any position on the path using the local version of the law (Equation 6). Here C denotes the curve, $W(s)$ the width of the tunnel at the position s and τ an empirically determined constant.

$$MT = a + b\int_C \frac{ds}{W(s)} \tag{5}$$

$$v(s) = \frac{W(s)}{\tau} \tag{6}$$

In theory, this law should be scale-invariant. This means that scaling down an experiment should not influence the movement time. Accot and Zhai showed that this is not the case [3]. Similarly to Fitts' Law, the Steering Law gets less precise when massively scaling the experiment up or down. This is due to the fact that the motor joint combination shifts or the control precision becomes the limiting factor. Still, scaling has much less influence than changing the index of difficulty.

Another model that predicted performance for trajectory-based movements has been developed by Cao et al. [8]. This research focused on pen-based interaction (especially gestures). Based on investigations of human handwriting, they built their gestures from straight lines, smooth curves and corners. They experimentally validated their models for each of the segments and created a model by adding up the partial times.

2.1.2 In Three Dimensions

Currently human computer interaction in three dimensions is becoming more and more common. There are several interfaces that can take input from the user in three dimensions. An example is the Kinect, a widely used input system developed by Microsoft. Current research tries to adapt the models to multiple dimensions.

The first attempt to model Fitts' Law in three dimensions was done by Murata and Iwase in 2001 [28]. They registered that the variance in movement time for a three-dimensional pointing task increased notably and could not be explained by current implementations of Fitts' law. To cater for the increased degrees of freedom, they performed their experiments with a variety of 8 approache angles θ, using two-dimensional circles as targets that were attached in the direction of the angles, in front of the participant. For the index of difficulty they derived Equation 7. Here D and W again denote target distance and width.

$$ID = \log_2\left(\frac{D}{W} + 1\right) + csin\theta \tag{7}$$

The increased degrees of freedom in three-dimensional movement also allowed for new types of interaction. Instead of touching targets with a finger, Kopper et al. investigated human performance in distal pointing tasks [21]. Distal pointing means pointing an input device at an interactive display in a distance, a common example is pointing with a Nintendo Wii controller. The model that was derived by Kopper used angular measurements to form the new Equation 8 for the index of difficulty. In this equation α is the angular amplitude of movement, ω is the angular size of the target and k is a constant power factor. This equation implies that the angular sizes, instead of the linear sizes of targets influence the difficulty of a distal pointing task.

$$ID = \left(\log_2\left(\frac{\alpha}{\omega^k}+1\right)\right)^2 \tag{8}$$

A similar pointing task but using another part of the body was evaluated by Zhang et al. [40]. They created a model for dwell-based eye-pointing tasks. They highlighted the difference in the transmission path of the electric signals that control the muscles in the eye. In contrast to hand or body movements the spinal nerve system is not incorporated in this task, solely the cranial nerve system. Additionally stabilizing an eye cursor is practically impossible because of involuntary eye jitter. This lead to the definition of the index of difficulty as stated in Equation 9. D and W are again target distance and width, μ is an empirically determined constant that defines a minimal target size. In contrast to other definitions this is not scale-invariant, since scale-invariance is not a property of eye-controlled interfaces.

$$ID = \frac{e^{\lambda D}}{W-\mu} \tag{9}$$

Other types of movement can be modeled in three dimensions as well. In the original publication about the steering law, its application for three dimensional movement already foreshadowed as Accot and Zhai highlight 'moving in 3D worlds' as a possible task for the application of their model [1]. Although the paper itself concentrated on two-dimensional movements, extending it to three-dimensions could be done by extending the tunnel from two to three dimensions.

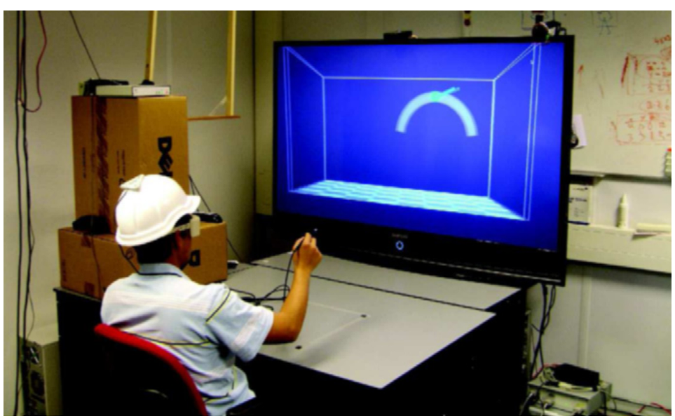


Fig. 3: Experimental setup for steering tasks in 3D. [22]

Liu. et al. revisited path steering in three dimensions in 2011 [22]. They not only used a six-degrees-of- freedom stylus, but combined it with a head tracked stereo display showing the task in 3D. The virtual representation of the pen had an offset from the physical device, thus motor space and visual space were separated, like in two-dimensional steering tasks with a mouse. With this setup, they modeled the task as moving a cursor ball through a semi-transparent tube of a fixed diameter. Figure 3 shows the experimental setup. By using a cursor ball, they decoupled the steering of position and orientation. As they modeled their task in a plane, it allowed them to rotate the task around the x and y axis by the angles α and β. Because of the fixed width of the tunnel, the index of difficulty according to the steering law (Equation 5) should have only been dependent on path width and length. Liu et al. found that this theory does not hold for three-dimensional tasks and showed that curvature of the path as well as the orientation of the path also contributed to the index of difficulty of a steering task. Based on their experiments they proposed a new model for steering tasks in a 3D environment using α, β, and the curvature of the task ρ as shown in Equation 10.

$$MT = e^{a+b\frac{L}{W}+c\rho+d\cos\alpha+e\cos\beta+f\sin\beta+g\cos 2\beta+h\sin 2\beta} \tag{10}$$

It is important to note that, although the task itself was designed in three dimensions, the described trajectory lay on a plane and was therefore two-dimensional. A model for arbitrary movements in three dimensions does not exist yet.

Using a similar setup as in the previous work Liu et al. also investigated the performance of object pursuit tasks in a VR environment [23]. In this interaction task, the users were required to track a ball, moving with a uniform velocity, using a tracker they hold in their hands. The ball stopped when the cursor of participant moved outside it and started again once the participant had focused it again. They proposed that these kind of tasks can be broken down to a tracking phase and a correction phase, that happened when the user loses focus on the ball. They proposed and validated Equation 11 as an prediction for the movement time, where L is the length of the path the target traveled, v is it's speed and W its width. The model has only been evaluated for targets moving in linear or circular paths.

$$MT = \frac{L}{v} + e^{a+b\frac{L}{W}+cv} \tag{11}$$

2.2 Error Rate

When carrying out studies about models for movement time one wonders whether the participants never miss targets and what happens to erroneous trials. Usually, when testing for movement time, the participants are requested to execute the experiment as fast as possible, but keep the error rate below a certain level (for example 4% for the original Fitts' Law experiments [15]). Experiments that do not fall below this error rate are filtered from the results. Additionally a lot of user interfaces are not used in agreement with movement time laws, because users might excercise extra care or hastiness. Knowing how many errors happen with such interfaces can be important, for example when designing buttons in safety critical applications where speed matters as well as safety.

In 2008 Wobbrock et al. highlighted that no research up to this point had extracted a predictive error model from Fitts' Law [36]. They showed that Fitts' Law mathematically implies an equation for pointing errors, not only a spread of hits. Their findings were based on work that shows that rapid aimed movements have a normal distribution around the target's center [16] and the fact that Fitts' Law applies when the error rate is at a certain level. Equation 12 shows the error model they derived, where erf is the Gauss error function and MT_e is the movement time of the user. W and D denote the target width and distance. They showed experimentally that this model held over a range of target distances, sizes and movement times.

$$\begin{gathered} P(E) = 1 - erf\left(\frac{2.066\frac{W}{D}\left(2^{\frac{MT_e-a}{b}}-1\right)}{\sqrt{2}}\right) \\ erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \end{gathered} \tag{12}$$

Similarily Guiard et al. proposed a new form of Fitts' Law, that describes the model as a time-error-tradeoff rather than the classical time-accuracy tradeoff [17]. Equation 13 shows their definition of error rate, where q and p represent adjustable coefficients. To verify their findings they used the data provided by Fitts in the original experiments and performed additional experiments.

$$P(E) = \frac{MT}{q}^{\frac{1}{p}} \tag{13}$$

As models got more complex over time Wobbrock et al. also adapted their error model to pointing tasks in two dimensions [37]. Based on their previous research [36] they extended their model to use bivariate normal distribution for the endpoints. The equation for the error rate therefore was no different from Equation 12. They validated their findings in a study with 21 participants and showed that their error model held for two-dimensional tasks. Their results also concluded that bivariate normal distribution produced a similar match as the univariate distribution.

An improvement from simply predicting error rates for pointing tasks was presented by Banovic et al. [6]. They showed how the cost of errors C_r and the time needed to recover from an error MT_C influenced the completion time for a two-dimensional pointing task. Equation 12 from Wobbrock et al. was used to estimate the probability $P(E)$ of an error and the Shannon formulation of Fitts' Law was used for the prediction of the movement time MT. They estimated the expected completion time as seen in Equation 14. This equation included that the users adapt the way that they execute target-directed pointing when time-based penalties for errors are introduced.

$$CT = (1-P(E))MT + P(E)(MT + C_r + MT_C) \tag{14}$$

Other research focused on the deviations from the optimal path for the Steering Law. Zhou et al. provided a model in [41] that predicted standard deviation from the optimal trajectory for a fixed movement time. To do this they proposed an experiment where movement time MT, the tunnel width W and length L were defined for the participants. Their hypothesis was that similarly to pointing tasks, the standard deviation is related to average movement speed and the tunnel width as seen in Equation 15. Note that they were only defining standard deviation, not out-of-path movement which would be the equivalent to errors in pointing tasks, although they measure the out-of-path movements in their experiments. They also did not show how to extract out-of-path movements from the standard deviation.

$$SD = a + bW + c\left(\frac{L}{MT}\right) \tag{15}$$

2.3 Throughput

Throughput is another aspect of the user that can be modeled in human computer interaction. Throughput relates to the amount of information that can be transmitted between a human and the computer when using a certain input device for a specific task. Throughput is also called index of performance (IP) and is measured in $\frac{Bits}{s}$.

In respect to Fitts' Law there exist two definitions of throughput: Equations 17 and 16. Equation 17 is the theoretical definition of throughput defined by the Shannon Formulation. Equation 16 is the definition that is widely used in practice. Both allow estimating and comparing performance for certain limbs as done by Balakrishnan et al. [5], who computed throughputs for several parts of the human upper limb in a pointing task. Equation 16 has even been used in ISO 9241-9 to provide a standard for input device evaluation.

$$TP = \frac{ID}{MT} \tag{16}$$

$$TP = \frac{1}{b} = \frac{ID}{MT - a} \tag{17}$$

Zhai et al. highlighted that Equation 17 as the choice of IP to quantify motor performance was suboptimal [39]. Their goal was to identify the minimum number of parameters needed to characterize input performance for pointing tasks and concluded that the intercept a needed to be used as well as the slope b. They highlighted that not only the slope b of the regression line influences performance, but the intercept a does as well. In previous research the intercept a was attributed to different interactions, for example the initial adjustment of the hand on the mouse or the time needed for button presses. Zhai et al. identified some additional non-information aspects of pointing that were possible sources of a non-zero a, for example reaction or activation processes of the human motor system.

Olafsdottir et al. built on this research and discussed further whether the intercept a needed to be included in the calculations for throughput as seen in Equation 17. Additionally they showed that the order in which aggregation and throughput calculation is done influenced the result for throughput [30]. This influence is called Jensen's Inequality. They showed that the traditional way of calculating throughput through Equation 16 failed invariance tests and since research was using different orders of calculation and aggregation, the resulting values for throughput were rarely comparable. They proposed a systematic investigation of these issues and believed that the Jensen's Inequality is a general methodological issue.

Aside from extracting throughput from the index of difficulty and the movement time of participants there were also other approaches that focused on calculating throughput from other sources. For information channels throughput is defined as the rate of successful message delivery, meaning the message has been reproduced and received exactly as planned. Similarly throughput in human computer interaction can be defined as the human ability to exactly reproduce a movement [33]. As the throughput of a motion can be modified intentionally by its performer, most of these works cope with information capacity. Information capacity is the maximum throughput the user could have transmitted with.

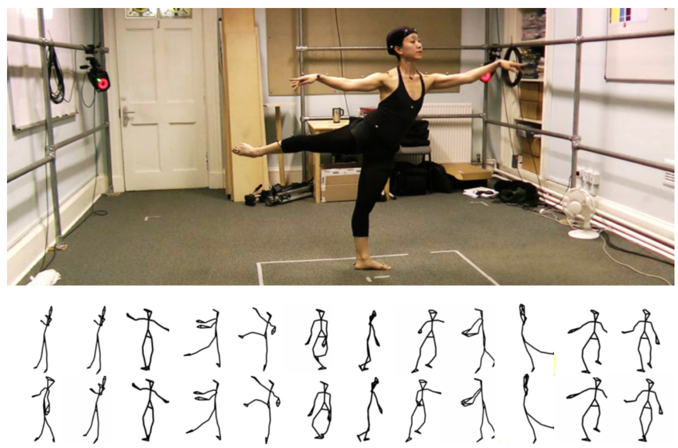


Fig. 4: Recording the ballet dataset, a sequence and its repetition. [33]

This theoretical basis was used by Oulasvirta et al. to extract the information capacity of full-body movements using motion-capture recordings [33]. They tried to introduce a measure of the amount of controlled information that was included in multiple executions of the same movement. As identically executing a movement at exactly the same speed and using the exact same poses is practically impossible, they used time warping to do temporal alignment and dimension reduction to filter dimensions without information. To calculate the maximum capacity of fully body movements, they conducted a study with a professional ballet teacher, who was asked to repeat a dance as exact as possible as seen in Figure 4. From this data the information capacity of the human-computer channel was calculated. The results stood for themselves as the calculated information capacity would have required the dancer to be able to perform all other movements in the motor space as well. Additionally they compared their calculations of throughput with calculations from Fitts' Law using a aimed-movement task and came to the conclusion that the two throughput metrics cannot be used to compare the performance for the same task. Nevertheless they introduced a new metric based on the shape and reproducibility of a trajectory.

Other approaches to calculating information capacity were done in the field of biomedical engineering. Instead of using recordings of motions, they used biomechanical models of the human motor and nervous system to estimate the information capacity of limbs. One of these works looked at the information capacity of the thumb and index finger in communication [25]. This allowed to optimize hand gestures, for example for American Sign Language (ASL). To extract the information capacity, they estimated how many poses can be reliably communicated from the brain to the hand, without being altered by noises occurring in the human nervous system. They concluded that the brain can deliver a maximum of ten bits of information to the thumb and seven bits of information to the index finger. Based on these results they estimate the information capacity of the hand at 150 bits per second.

2.4 Perception

Former research has commonly modeled user perception in a descriptive manner. Based on the target of finding which arrangement of menu items was the easiest for users, Card researched what perceptual models were employed by humans when they searched through linear menus [9]. He found that instead of searching linearly through the menu, the user rapidly focused random parts of the screen with fast eye movements called saccades. From this he derived a model for the mean time to detect a targeted menu item that is shown in Equation 18, where p_s is the probability of finding an item with a single saccade and T is the fixation time after a saccade. Both of these variables were experimentally extracted.

$$t_s = \frac{T}{p_s} \tag{18}$$

In more recent research, cognitive architectures like ACT-R or EPIC were used to simulate cognitive and motor processes. Hornof and Halverson used EPIC to model search in hierarchical menus [20]. They implemented a reproduction of the task environment, cognitive strategies and the perceptual features of the menu items in the framework and let it predict eye movement and search time. Although the results were satisfying according to the authors, there was not enough data released to reproduce their experiments.

In [18] Halverson and Hornof extend their research to be no longer restricted to hierarchical menus. They proposed a new minimal model, which only used three assumptions about search strategies: (1) Eye movements tend to go to nearby objects, (2) fixated objects are not always identified and (3) eye movements start after the fixated objects are identified. They again implemented these assumptions in the EPIC framework and validated their new model with an experiment. This minimal model adapted to the task better than the previous model.

However, not only the performance of perception can be modeled, the perception of performance can be predicted as well [29]. Nicosia et al. extracted a model for the perception of performance for an aimed movement task. They showed how changing the index of difficulty for such a task influenced the perception of the users performance and extracted a mathematical model that predicted whether the user registered a performance difference in his or her task. The probability of noticing a difference of x in ID is defined in Equation 19.

$$p(x) = \frac{1}{1+e^{-(ax+b)}} \tag{19}$$

3 Creating Models

Following this review of predictive models, we want to show how models are extracted and adapted from data collected from studies. Usually this is done by collecting the data, creating a theoretical model for this data and then using linear or non-linear regression to validate it by fitting a curve [24]. Outliers, meaning data that differs significantly from the rest, are removed in the process. The regression is used to extract the parameters of the model from the experimental data. For Fitts' Law these parameters are a and b. Regression is chosen because it incorporates many goals of the researchers: It is computed rapidly and it allows to easily adapt model fit by introducing additional parameters.

Several frameworks exist to help with this process. They allow to investigate the relationships between dimensions and visualize the distribution and key figures of the independent variables [27]. This assists in extracting the theoretical model, because it allows the researcher to see which variables correlate with the measured data. Some of these frameworks are even specialized on regression models for human computer interaction, for example the 'Movement Time Evaluator' by Scheldbauer [34]. These frameworks also help the researchers to create and record experiments and graphically explore the recordings.

Another approach is to generate the regression model automatically. Oulasvirta presented an algorithm that iteratively builds a model from collected data, based on observations that have been made in human computer interaction: (1) Particular operations on terms are favored, (2) the number of predictors is small, (3) a linear prediction can serve as an initialization for search and they are typically based on averaged point data [31]. Extracted models take the form as seen in Equation 20.

$$Y_i = \sum_{j=1}^{M} \beta_i f_j(x_i) + \varepsilon_j \tag{20}$$

In his paper he calculated several best fitting models for historical studies employed by other researchers in the human-computer-interaction field. He showed that the automatic model extraction yielded better results (meaning a significantly higher correlation coefficient) than some of the models proposed by the original authors.

4 Applications

Applications for the predictive models are manifold. At first Fitts' Law and its derivatives were mostly used to compare the efficiency of limbs of the human body [5]. It showed that the hand has significantly better throughput and error rate that other limbs. A similar study, but focused on different input devices was executed with the Steering Law for trajectorial tasks [2]. The study showed that mouse and a tablet with stylus were superior input systems for trajectorial tasks compared to a trackball, touchpad and trackpoint. As evaluations like these were the initial purpose of the laws, since they were quantifying the accuracy of human movement, these kind of applications were diverse. A similar study was done by Cockburn et al. to compare several methods of helping the user to acquire small targets by mouse [12]. He compared the expanding, sticky and goal-crossing targets metaphors that were employed to aid the user to select small targets. He showed that all of these techniques improved the task completion time and that Fitts' Law could be employed for all of them.

Fig. 5: Two optimized keyboard layouts. [14]

In more recent research, Fitts' Law was often used to optimize the movement times in user interfaces. Dunlop et al. proposed a method to optimize on-screen touch keyboards based on three goals [14]: minimizing finger travel distance by moving keys using Fitts' Law, keeping the keyboard close to a standard QWERTY layout and introducing ambiguity to improve automatic error correction through spell checks. Such optimizations can be seen in Figure 5. They showed that the users performance with such a optimized layout can approximate the performance with an QWERTY keyboard without the user having as much familiarity with the new layout.

Oulasvirta et al. also optimized touchscreen keyboards, with their focus lying on split keyboards for two thumb entry [32]. They used a

Table 1: Model Overview: This table shows that most of the research is focused on two-dimensional aimed pointing movements

Researcher	Year & Publication	Task	Prediction	Predictors	Notes
Fitts	1954 [15]	1D Aimed Pointing	Movement Time, Throughput	D, W	
McKenzie	1992 [24]	2D Aimed Pointing	Movement Time, Throughput	D, W, H, θ	
Accot	2003 [4]	2D Aimed Pointing	Movement Time, Throughput	D, W, H	
Grossmann	2005 [16]	2D Aimed Pointing	Movement Time, Throughput	$P(h)$	
Bi	2013 [7]	2D Aimed Pointing	Movement Time, Throughput	D, σ, σ_a	Touch input only
Banovic	2013 [6]	2D Aimed Pointing	Movement Time including Errors	$P(E), MT, MT_C$	
Wobbrock	2008 [36]	2D Aimed Pointing	Error Rate	W, D	
Murata	2001 [28]	3D Aimed Pointing	Movement Time, Throughput	D, W, θ	All targets in a plane
Accot	1997 [1]	2D Trajectory Based	Movement Time, Throughput	$C, W(s)$	
Cao	2007 [8]	2D Trajectory Based	Movement Time	Building Blocks of a curve	
Zhou	2009 [41]	2D Trajectory Based	Standard Deviation from Trajectory	W, L, MT	
Liu	2011 [22]	3D Trajectory Based	Movement Time, Throughput	$\alpha, \beta, \rho, L, W$	Trajectory in a plane
Oulasvirta	2013 [33]	2D/3D Trajectory Based	Throughput		
Kopper	2010 [21]	3D Distal Pointing	Movement Time	α, ω	
Zhang	2010 [40]	3D Distal Pointing	Movement Time, Throughput	D, W	Eye, All targets in a plane
Liu	2011 [22]	3D Object Pursuit	Movement Time, Throughput	L, W, v	
Mao	2009 [25]	Hand Gestures	Throughput		

Common Predictors. D: distance to target, W: target/tunnel width, H: target height, θ: approach angle, C: trajectory, L: length of trajectory, v: velocity, MT: movement time

predictive model for movement time between the keys as well, adapted to two-thumb input and based on Fitts' Law. They found that alternation between thumbs and same-side tapping with the dominant hand were the fastest kinds of typing. The resulting layout proved to be superior to a full-width QWERTY keyboard by about 30%.

Another focus of research was the performance of the user with menus employed in current systems. Cockburn et al. use a combination of Fitts' Law for movement times and the Hick-Hyman Law for decision times [13]. They compared several menu layouts and showed that hierarchically arranging the items is superior to flat arranging.

5 Discussion

Now, after we have highlighted several models and their applications, we want to discuss the state of research of models in human computer interaction. For this purpose we created Table 1 that gives an overview over all models included in this paper that somehow model human movement.

All except of a few models (Cao, Liu, Mao and Oulasvirta) work with Shannon's speed-accuracy tradeoff formula or are derived from it. This shows that the speed-accuracy tradeoff is a proven and flexible concept that can be adapted to a lot of situations. Usually only a redefinition of the index of difficulty is necessary to model a new task.

Of all tasks, the one the most researched is aimed pointing in two dimensions (Accot, Banovic, Bi, Grossmann, McKenzie, Wobbrock). This probably originates from Fitts' Law that also handled rapid-aimed movements. As two-dimensional pointing is very common in current systems, there are also the most applications of human models in this area. Trajectory-based movements have also been researched well in two-dimensions (Accot, Cao, Zhou). Their applications are more sparse currently, probably to the cause that they are not as common. Extending the movements to three dimensions has not been very successful yet. When considering the proposed models it is necessary to note that they executed their experiments in very restrictive ways. In these models the targets or trajectories tend to lie in a plane, which is not always a realistic task (Liu, Murata). A model for arbitrary movements, either trajectoral or aimed is not existent yet. This needs to be taken into account when working with models for three-dimensional movements and could be the focus of future research.

New types of tasks emerged recently as the research in the field of three-dimensional models intensified. One example for them is the distal pointing task. Zhang unfortunately modeled this type of task as a two-dimensional task, since they keep all targets in a plain and the distance from this plain was not incorporated into the model. Kopper showed how this task can be better mapped to a 2D task, by using angular sizes and distances. These can be calculated from the three-dimensional coordinates of the target, but still allowed the task to be considered three-dimensional.

In perception the focus lies on building models using the existing architectures like ACT-R or EPIC. While this allows for very fine grained control over the model (for example eye movements can be predicted), it also introduces some complexity. In contrast to the simple models for movement time, models created with cognitive architectures can rarely be expressed by a simple formula. This makes the validation of these models very complicated for other researchers. It is up to discussion whether an easier model for these kind of tasks exists.

In conclusion the review of user models in this paper has shown that most research focuses on the approach that Fitts' Law introduced. This leads to a relatively slow rate of innovation in this field as new models are rather adoptions than built from the ground up. In more recent years this has changed, because there have been more efforts to quantify user interactions in HCI. The research of Liu et al. indicates that Fitts' Law could be at its boundaries for three-dimensional tasks. To really advance the field, future research should focus on more diverse modeling to be able to describe more complex tasks, for example in ubiquitous computing or virtual reality.

6 Conclusion

In this paper we have presented several predictive models for human-computer-interaction tasks. We have introduced models for aimed pointing and trajectory-based tasks in two and three dimensions. Other modeled tasks include distal pointing and object pursuit. We have also presented ways to determine the information capacity of the human-computer channel. We discussed how these models are created and can be applied to optimize user interfaces.

References

[1] J. Accot and S. Zhai. Beyond fitts' law: models for trajectory-based hci tasks. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 295–302. ACM, 1997.

[2] J. Accot and S. Zhai. Performance evaluation of input devices in trajectory-based tasks: an application of the steering law. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 466–472. ACM, 1999.

[3] J. Accot and S. Zhai. Scale effects in steering law tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1–8. ACM, 2001.

[4] J. Accot and S. Zhai. Refining fitts' law models for bivariate pointing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 193–200. ACM, 2003.

[5] R. Balakrishnan and I. S. MacKenzie. Performance differences in the fingers, wrist, and forearm in computer input control. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 303–310. ACM, 1997.

[6] N. Banovic, T. Grossman, and G. Fitzmaurice. The effect of time-based cost of error in target-directed pointing tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1373–1382. ACM, 2013.

[7] X. Bi, Y. Li, and S. Zhai. Ffitts law: modeling finger touch with fitts' law. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1363–1372. ACM, 2013.

[8] X. Cao and S. Zhai. Modeling human performance of pen stroke gestures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1495–1504. ACM, 2007.

[9] S. K. Card. User perceptual mechanisms in the search of computer command menus. In *Proceedings of the 1982 conference on Human factors in computing systems*, pages 190–196. ACM, 1982.

[10] S. K. Card, A. Newell, and T. P. Moran. The psychology of human-computer interaction. 1983.

[11] O. Chapuis and P. Dragicevic. Effects of motor scale, visual scale, and quantization on small target acquisition difficulty. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(3):13, 2011.

[12] A. Cockburn and A. Firth. Improving the acquisition of small targets. In *People and Computers XVIIDesigning for Society*, pages 181–196. Springer, 2004.

[13] A. Cockburn, C. Gutwin, and S. Greenberg. A predictive model of menu performance. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 627–636. ACM, 2007.

[14] M. Dunlop and J. Levine. Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2669–2678. ACM, 2012.

[15] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.

[16] T. Grossman and R. Balakrishnan. A probabilistic approach to modeling two-dimensional pointing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(3):435–459, 2005.

[17] Y. Guiard, H. B. Olafsdottir, and S. T. Perrault. Fitt's law as an explicit time/error trade-off. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1619–1628. ACM, 2011.

[18] T. Halverson and A. J. Hornof. A minimal model for predicting visual search in human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 431–434. ACM, 2007.

[19] W. E. Hick. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1):11–26, 1952.

[20] A. J. Hornof and T. Halverson. Cognitive strategies and eye movements for searching hierarchical computer displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256. ACM, 2003.

[21] R. Kopper, D. A. Bowman, M. G. Silva, and R. P. McMahan. A human motor behavior model for distal pointing tasks. *International Journal of Human-Computer Studies*, 68(10):603–615, 2010.

[22] L. Liu, J.-B. Martens, and R. Van Liere. Revisiting path steering for 3d manipulation tasks. *International journal of human-computer studies*, 69(3):170–181, 2011.

[23] L. Liu and R. van Liere. Modeling object pursuit for 3d interactive tasks in virtual reality. In *Virtual Reality Conference (VR), 2011 IEEE*, pages 3–10. IEEE, 2011.

[24] I. S. MacKenzie and W. Buxton. Extending fitts' law to two-dimensional tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 219–226. ACM, 1992.

[25] Z.-H. Mao, H.-N. Lee, R. J. Sclabassi, and M. Sun. Information capacity of the thumb and the index finger in communication. *Biomedical Engineering, IEEE Transactions on*, 56(5):1535–1545, 2009.

[26] D. E. Meyer, R. A. Abrams, S. Kornblum, C. E. Wright, and J. Keith Smith. Optimality in human motor performance: ideal control of rapid aimed movements. *Psychological review*, 95(3):340, 1988.

[27] T. Muhlbacher and H. Piringer. A partition-based framework for building and validating regression models. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):1962–1971, 2013.

[28] A. Murata and H. Iwase. Extending fitts' law to a three-dimensional pointing task. *Human movement science*, 20(6):791–805, 2001.

[29] M. Nicosia, A. Oulasvirta, and P. O. Kristensson. Modeling the perception of user performance. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1747–1756. ACM, 2014.

[30] H. B. Olafsdottir, Y. Guiard, O. Rioul, and S. T. Perrault. A new test of throughput invariance in fitts' law: role of the intercept and of jensen's inequality. In *Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers*, pages 119–126. British Computer Society, 2012.

[31] A. Oulasvirta. Automated nonlinear regression modeling for hci. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3899–3902. ACM, 2014.

[32] A. Oulasvirta, A. Reichel, W. Li, Y. Zhang, M. Bachynskyi, K. Vertanen, and P. O. Kristensson. Improving two-thumb text entry on touchscreen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2765–2774. ACM, 2013.

[33] A. Oulasvirta, T. Roos, A. Modig, and L. Leppänen. Information capacity of full-body movements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1289–1298. ACM, 2013.

[34] M. J. Schedlbauer. An extensible platform for the interactive exploration of fitts' law and related movement time models. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*, pages 2633–2638. ACM, 2007.

[35] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[36] J. O. Wobbrock, E. Cutrell, S. Harada, and I. S. MacKenzie. An error model for pointing based on fitts' law. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1613–1622. ACM, 2008.

[37] J. O. Wobbrock, A. Jansen, and K. Shinohara. Modeling and predicting pointing errors in two dimensions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1653–1656. ACM, 2011.

[38] R. S. Woodworth. Accuracy of voluntary movement. *The Psychological Review: Monograph Supplements*, 3(3):i, 1899.

[39] S. Zhai. Characterizing computer input with fitts law parametersthe information and non-information aspects of pointing. *International Journal of Human-Computer Studies*, 61(6):791–809, 2004.

[40] X. Zhang, X. Ren, and H. Zha. Modeling dwell-based eye pointing target acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2083–2092. ACM, 2010.

[41] X. Zhou, X. Cao, and X. Ren. Speed-accuracy tradeoff in trajectory-based tasks with temporal constraint. In *Human-Computer Interaction–INTERACT 2009*, pages 906–919. Springer, 2009.

Simplification of Human-Robot Interfaces

Lukas Ziegler

Abstract— Advances in robot technology have increased the presence of robots in our daily lives and the demand on human-robot interaction (HRI). The efficient and intuitive interaction between humans and robots has evolved to an essential part in the construction of human-robot interfaces. In this study we observe what has been accomplished in previous works and evaluate which metrics and common approaches have been found so far, allowing the reader to build upon the expertise of previous studies. The goal is to summarize principles for designing efficient and easy-to-use user interfaces, to give an overview of recent developments and to analyze current trends. These findings can be used for constructing simpler and user-friendlier interfaces, leading to a higher overall human-robot performance. Possible use cases are amongst others the remote camera control.

Index Terms—HRI, interfaces, situation awareness, usability, simplification, metrics, guidelines

1 Introduction

Already since the late 20th century robots became more present in our daily lives, and in recent years the direct interaction between humans and robots also became more omnipresent. This trend towards a collaborative team-work of humans and robots is the foundation for this paper. The research field for human-robot collaboration is relatively interdisciplinary, the focus in the following will be on the research area of human-robot interaction (HRI).

The first public discussion of problems in HRI was in 1940 by Isaac Asimov [3], in which he stated the first three laws of robotics: 1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2) A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law. 3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. These laws were published, although no robots existed yet at the time. The vision was a safe interaction between humans and robots. In the years since these laws were created, robots were mainly used for industrial and military use cases [19]. Of importance was the automation of complex tasks, for which humans would require more time. Industrial robots used to be locked away from direct human interaction, allowing them to execute their tasks at a higher speed and precision, as humans ever could.

Due to recent advances in technology, the use of robots to extend and complement human capabilities increased. In various areas in our society robots became more present and are gaining a higher impact on our daily lives. Robot vacuum cleaners are offered in discounters, complex surgeries are run with the aid of robots, and numerous research labs are working on robots to work alongside humans, for example in kitchen environments [23]. While well established car manufacturers, like Volvo, Audi, BMW, and Toyota, are working on semi-autonomous driving cars, Google is already experimenting with autonomous self-driving cars [25, 21].

Since robots are evolving to more omnipresent machines, the demands on user-friendliness also increase. Back in the days when robots with human interaction were mainly used for military purposes or for urban search and rescue (USAR), the quality of the user interface wasn't as important as now. All operators were specially trained before using the robots and a well-designed prototype of an interface was usually enough. However we are currently at a stage where robots are about to make the transition from being well-designed prototypes to becoming solid real-world products for the consumer market. The competition on the consumer market is much harder and the demands on usability will ever increase.

Fig. 1. The PdaDriver interface [12] is only one example of early-stage interface prototypes in HRI. Ways of simplification and improvement will be discussed in this paper.

The acceptance of robots is evolving from being prototypes (see *figure 1*) to becoming products making a profound impact on people's lives. "These will be robots that help us do things better, faster, and with greater knowledge about the world around us" [14]. Easy-to-use interfaces are of fundamental importance to the development of robots with an impact on our private lives.

To only name one example of robots becoming more present in our lives: Quadrocopters (unmanned, remote-controlled flying drones) are already being used for monitoring complex situations by firefighters, but also for commercial and leisure activities. In recent years quadrocopters were being used for recording images and movies in large and medium movie productions. Because of the ever-decreasing price, quadrocopters are now also being used by photographers working for news agencies to increase image quality and to capture events from completely new perspectives [26].

In areas such as search and rescue a good awareness is crucial for the remote operation of robots, however it is hard to attain a good situation awareness (SA) for remote controlled robots [27]. Therefore usability aspects of the remote interface play a crucial role.

A study from Yanco et al. in 2004 [27] showed that 30% of the interaction time was spent with acquiring or maintaining situational awareness. *Situation awareness* refers to the human's comprehension of the robot's environment and state. A higher situation awareness implies a faster completion of a given task. This is also the reason why we focus on finding guidelines for the creation of simple and easy to use interfaces. User-optimized interfaces allow the user to acquire

- *Lukas Ziegler is studying Media Informatics at the University of Munich, Germany, E-mail: lukas.ziegler@campus.lmu.de*
- *This research paper was written for the Media Informatics Advanced Seminar, in June 2014*

a higher situation awareness of the remote environment, leading to a faster completion time and a more efficient usage of the robot. This shows the demand and importance to further work on and improve interfaces in HRI.

In addition to this trend, the recent explosion of touch interfaces and the herewith combined development of simpler and easier-to-use interfaces already began to show its influence on this area of HRI. Users desire simpler interfaces with more natural means of interaction, based on touch-, gesture- or voice-based input technologies. In related work from Keyes et al. [17] this transition from keyboard- and joystick- based input devices towards touch-based user interfaces is described. Their study showed that using touch-based interfaces does not degrade the performance, but actually slightly improves the overall performance.

The rest of this paper is organized as follows: Section 2 covers related work and section 3 states human issues having an impact on HRI. Section 4 introduces metrics to measure the performance and efficiency of human-robot interfaces, which will be the base for deriving principles for simplification in Section 5. Section 6 applies selected metrics and guidelines to a real world example. In Section 7 recent trends in the interface design of HRI will briefly be discussed. A conclusion complements this paper.

2 Related Work

To get a basic understanding of HRI and the related aspects of usability, an overview of previous works relevant to this paper is given. The publication from Chen et al. [6] gives an extensive overview of the whole research area of HRI, the paper form Jaschkowitz [15] originated from the last years Media Informatics Advanced Seminar and introduces the reader in a structured way to navigational tasks.

For an introduction to HRI with a focus on usability and the simplification of interfaces, the papers from Yanco, Drury and Baker are of interest. All of them worked on related topics already around ten years ago. The publication form Yanco et al. [27] in 2004 deals with the human issues to attain SA and introduced metrics to measure SA implicitly and explicitly. Baker et al. [4] analyze the characteristics human operators show when using human-robot interfaces. Their study has shown that users rely heavily upon the video stream, ”to the exclusion of all other information on the interface”[4]. In the paper from Drury et al. [9] methods from human-computer interaction (HCI) and USAR are applied to HRI. To help their analysis, they followed a very fine-grained definition of awareness and describe a taxonomy of HRI-related characteristics.

For good foundation papers to measure the performance of human-robot interfaces, the publications from Goodrich, Olsen and Fong are relevant. Goodrich and Olsen [13, 20] introduced metrics to measure the efficiency of interfaces. Fong et al. [11] try to find common metrics for HRI and discuss the need for a toolkit for HRI metrics. In spite of the age of these papers, they are still relevant to current analysis. For metrics measuring quantitative team performance, see Singer et al. [22].

Micire's PhD thesis [18] and the chapter from Keyes et al. [17] are ideal to analyse the evolution of joystick-based interfaces (first introduced in [4], see *figure 3*) to interfaces supporting multi-touch interactions. The different revisions and aspects influencing the design decisions can be seen.

3 Human Issues

Understanding the influence human issues have on the way we perceive and interact with remote robot interfaces is of fundamental importance. In the following we will characterize and quantify a few of the most relevant human issues influencing the remote control of robots.

A variety of human factors influence the way we perceive interfaces for remote controlling a robot. The broadest categorization of human issues is on the level of remote manipulation and remote perception, as described in [6]. *Remote manipulation* is a fundamental part of the robot operator's task. It describes tasks required for the manipulation and navigation in the remote environment. The second aspect, the *remote perception*, focuses on the human side of the remote perception and is essential for effective teleoperation. The challenge in remote perception is mainly due to the decoupling of the remote environment from the physical (human-centered) environment.

To start off with, we will first give a brief overview of some widespread human issues, and thereafter limit our focus on one issue, being most relevant to this paper. The five human issues found in literature by Chen et al. [6] include: *motor skills* (intentional movement to perform a goal-orientated task), *obstacle detection* (the cognition of obstacles in the remote environment), building *mental models* of the remote environment, *distance estimation*, and *situation awareness* (awareness of what is happening in the surrounding).

For this paper human issues based on remote perception and situation awareness (SA) are of higher importance, since the focus lies upon the evaluation and simplification of interfaces. Of course multiple factors play a role in this domain, however the degree of remote perception and the attained situation awareness are good indicators for determining the quality of human-robot interfaces. The other human issues listed above play a subordinate role. A poor interaction with the interface influences the attainable SA.

The most widely accepted definition of SA was introduced by Endsley in 1998. Endsley distinguished between three levels of SA: ”The perception of elements in the environment within a volume of time and space (level 1 SA), the comprehension of their meaning (level 2 SA) and the projection of their status in the near future (level 3 SA)” [10]. For the purpose of HRI we will work with a more specific definition consisting of five different categories of awareness, introduced by Drury et al. [8], in the following cited directly form the source.

1. *Human-robot awareness*: the understanding that the humans have of the locations, identities, activities, status and surroundings of the robots. Further, the understanding of the certainty with which humans know the aforementioned information.
2. *Human-human awareness*: the understanding that the humans have of the locations, identities and activities of their fellow human collaborators.
3. *Robot-human awareness*: the robots' knowledge of the humans' commands needed to direct activities and any human-delineated constraints that may require command noncompliance or a modified course of action.
4. *Robot-robot awareness*: the knowledge that the robots have of the commands given to them, if any, by other robots, the tactical plans of the other robots, and the robot-to-robot coordination necessary to dynamically reallocate tasks among robots if necessary.
5. *The humans' overall awareness*: the humans' understanding of the overall goals of the joint human-robot activities and the measurement of the moment-by-moment progress obtained against the goals

As mentioned by Singer and Akin [22], SA can refer to three different perspectives of awareness: ”that of a human supervisor referring to knowledge of overall mission operations, that of a robot control operator referring to knowledge of the robot's immediate environment and obstacles, and of the robot's awareness of its own environment”. Although all three perspectives cohere, we will discuss situation awareness as the knowledge an operator has from the robots' immediate environment. From these five categories of SA, the *human-robot awareness* and the *humans overall awareness* are most relevant to the analysis in the following, because their focus is on the operators' awareness, in contrast to criterias 2 to 4. This grouping of awareness allows us to better evaluate criterias for designing user-friendlier interfaces.

4 Metrics

In this section we will deal with metrics used to measure the performance and efficiency of interfaces. Only when we know which element of the interface causes a problem, then we can strive to find solutions and work on improvements. The goal is to reduce cognitive load for acquiring situation awareness and thereby to enhance the ability to navigate more precisely in remote environments.

Goodrich and Olsen [13] introduced five metrics to measure the efficiency of interfaces, being summed up and described in the following. The metrics serve as a foundation for the simplification of interfaces based on empirical data. The first two metrics are directly measurable, while metrics three to five can be derived from the previous.

1. *Neglect time (NT)* is "a measure of how the robot's current task effectiveness declines over time when the robot is neglected by the user" [13].

2. *Interaction time (IT)* refers to "the amount of time required before performance rises from threshold to peak performance" [13], after having focused on a secondary task.

3. *Robot attention demand (RAD)* resembles the amount of time a robot demands in total and is given by $RAD = \frac{IT}{NT+IT}$.

4. *Free time (FT)* measures" how much time is left over for other tasks and is given by $FT = 1 - RAD$" [13].

5. *Fan out (FO)* is a measure for how many secondary tasks "can be effectively handled by a human" [13]. One use case would the operating of multiple robots in parallel by one human operator. FO depends on RAD and has an upper bound of $FO \leq \frac{1}{RAD}$.

Another way to measure the performance of a robot was summarized by Singer and Akin [22]. This approach first measures task specific metrics and subsequently measures the performance of the team (human + robot) as a whole. The overall ranking states how well the main objective was fulfilled. In literature numerous metrics to measure the interaction performance are mentioned. To only list a few of those metrics, performance can be measured by the number of collisions, the mean time between failures/collisions, the robot's resource usage or by the task dependent reliability or overall reliability. Kannan [16] suggests calculating the mean time between interventions (MTBI) or by the resiliency to failure. Schreckenghost introduced a work efficiency index (WEI), measred using the NASA-TLX model (only useful for a priori analysis). For an overview and discussion of these and further metrics, refer to "A Survey of Quantitative Team Performance Metrics for Human-Robot Collaboration" by Singer and Akin [22]. For most surveys the five metrics from Goodrich and Olsen are sufficient. In the next step we will carry on and use these metrics to construct and evaluate good interfaces for human-robot interaction.

5 Simplification of Interfaces

In the following, we will look at common principles being used to improve and simplify human-robot interfaces. At first findings from individual studies are presented and in the second half a collection of best-practices for simplifying robot interfaces are gathered.

5.1 Analysis

Baker et. al [4] recommend making suggestions the user on which *autonomy mode* he should be using. First of all, it is important to differentiate between and implement different autonomy modes, allowing the user to decide by himself how much control and effort he wants to put into a specific task. The INEEL system [1] works with four autonomy modes: teleoperation, safe, shared and autonomous (see the far right side of *figure 2*). The teleoperation mode allows the operator to have full control of the robot, leading to the highest cognitive load for the user. In safe mode the operator still is in charge of the commands, but the robot will prevent making collisions. In shared mode the robot only takes directions, where to navigate to, and the robot will take care of getting there safely. The autonomous mode resembles a mode of full automation, where the operator does not have to interact with the interface. Differentiating between autonomy modes, does not only give the operator a more fine grained control of navigation, but also significantly reduces the cognitive load.

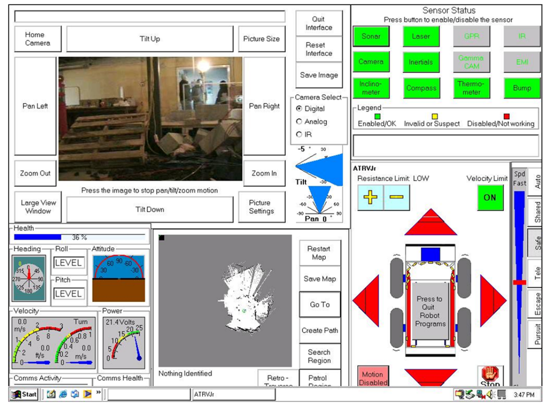


Fig. 2. The INEEL interface for their robotic system [5].

Furthermore the authors recommend simplifying and reducing the amount of visual clutter on the interface itself. The first step is to *differentiate between controls and indicators*. It is important for the user to know with which icons he can interact and what is only intended for the transmission of status information. Another step towards improving the interface can lead through providing *two types of panels*: one greatly simplified interface for most users (clean and reduced to a minimum) and one for expert users (with lots of information for debug purposes, typically used for development purposes). The goal for both interfaces, but especially for the end-user interface, is to reduce the amount of always visible status information. Most information can be hidden from the foreground and only displayed when it becomes relevant, such as uptime, battery level, geo coordinates of remote location, machine time or software version. These kind of system relevant information can be extracted and displayed when needed through priorized *system alerts*. Other measures for the simplification of human-robot interfaces include *sensor fusion* (combining sensor data into one unified representation) and *ranging information* (aligning information around the video stream).

When further extending upon these ideas, one can also *priorize and fuse* data from multiple sensors (sensor fusion). Displaying the data from all available sensors would normally lead to a cognitive overload and would not create any additional value. One example would be to use the distance sensors to detect obstacles, but to only visualize a warning, using visual overlays blended into the video stream, when the warning becomes relevant to the user. This notification can be coupled with the autonomy mode and only be triggered, when the interface is running in teleoperation mode, to prevent foreseeable collisions.

In a follow up publication from Drury and Yanco [9], they summarize the previous work from Baker et al. and state four high-level guidelines. The goal is to enhance awareness and reduce cognitive load on the user, while providing help in choosing the best autonomy mode for the robot and increasing the efficiency of every single task. All these aspects lead to an increased overall performance. When following these steps and focusing on reducing the visual clutter on the interface, an increased efficiency for HRI should be the effect.

1. enhance awareness

2. lower cognitive load

3. provide help in choosing robot modality (autonomy mode)

4. increase efficiency

From the work of Yanco et al. the following recommendations can be derived. After first evaluating map-centric interfaces they quickly saw the advantages of video-centric interfaces. This led to the strong

[1] An interface developed by the Idaho National Engineering and Environmental Laboratory

usage of video streams for acquiring SA. Their research has shown that "users rely heavily on the main video screen and very rarely notice other important information presented on the interface" [18]. Therefore they put all important information on or near the main video panel.

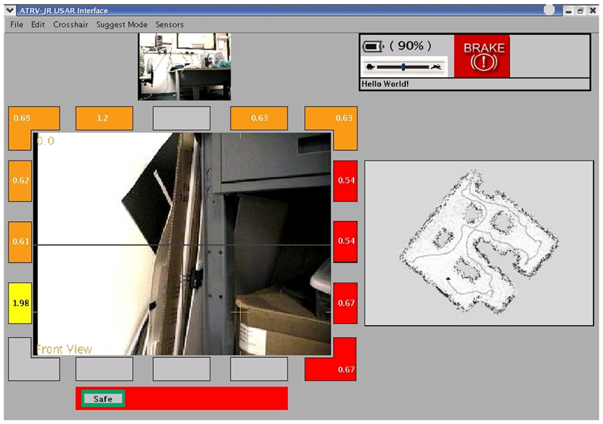


Fig. 3. The original UML USAR interface design [4] with an enlarged video stream and lots of the ranging information displayed on or near the video stream.

With the growing popularity of multi-touch interfaces this kind of input device was also analyzed for HRI. Their result is, that multi-touch interfaces do not degrade performance.

Minor weaknesses in the UI were eliminated through iterative design and evaluation cycles. Especially with the distance panel there was a need for improvement. In the first versions the users had problems interpreting the distance information correctly, since it still used a map-centric approach to display the distances. Their refinement involved the change to a more user-centric panel, which rotates when the operator pans the camera. This led to a reduced cognitive load and allows the users to match obstacles form the distance panel with the content from the video panel.

When following the recommendations from Baker et al., the simplified version of their interface gets close to the approach seen in *figure 5*. RoBrno's interface follows the suggestions of a clean and simplified user interface. All irrelevant information is hidden from the operator, the user can fully concentrate on the video stream and thus acquire a high situation awareness, in turn leading to less errors and a higher overall performance. In RoBrno's interface they also followed the suggestion from Baker et al. [4] to place all important information on or besides the video stream.

5.2 Guidelines

To complement the previous analysis, a collection of recommendations for designing a good remote robot interface can be found in the following. These guidelines originate from several referenced papers, amongst others [4, 9, 17]. In order to develop an easy-to-use and user-friendly interface, taking the following guidelines into consideration is of importance.

- provide consistency
- provide feedback
- use a clear and simple design
- ensure the interface helps to prevent and recover from errors
- follow best-practices and real-world conventions from other applications (touch or mobile interfaces)
- implement a forgiving interface, allow for reversible actions
- enable efficient operations with low demand on time
- map or at least monitor of where the robot is and has been
- fuse sensor information to lower the cognitive load
- minimize the use of multiple windows
- give spatial information about robot environment
- give suggestions for correct autonomy level
- show a reference frame to determine the relative position
- indicators of robots' health and state, including which camera is being used, the position(s) of the camera(s), traction information and pitch/roll indicators
- a view of the robots' body, in order to inspect for damage or entangled obstacles. Being able to see the robots' chassis enhances awareness
- enable an understanding of the robot's location in the environment
- show moment-by-moment progress of activities for a better understanding of the overall mission
- design the video screen as large as possible and locate the most important information on and around the video screen
- use a small crosshair to indicate the current direction of the camera
- add a rear-looking camera for reverse movements to enhance awareness
- design robots to come up with their own resolution plan when getting stuck or errors occur
- allow for varying levels of robot autonomy and suggest the best mode to the user, when it will improve the overall performance

6 Practical evaluation

The transfer of the theoretical concepts from previous sections to concrete interfaces from literature is the goal of this section. The focus will be on the University of Massachusetts Lowell's (UML) Urban Search and Rescue (USAR) interface introduced by Baker et al. in 2004 and subsequently improved over the years. In Micire's PhD thesis [18] there is a good evaluation of the subsequent improvements, with reasons why they improved certain elements.

The UML USAR interface (see *figure 3*) was introduced and developed by Scholtz, Yanco and Drury, and subsequently improved with several usability studies over the years. Their original interface design was based on the INEEL interface (see *figure 2*). The first improvements on designing their follow-up interface included:

1. The automatic recentering of the camera, after making pan and tilt operations.
2. Placing ranging information directly around the video stream, instead of creating a new panel.
3. Working with system alerts, in order to hide not always relevant information from the foreground.
4. Making suggestion for the best suited autonomy mode.
5. Including a map of the environment, to enhance SA. However working with maps turned out to be in need for improvement, as we discuss later on.
6. Allowing customizations of the interface.
7. Adding sensors to the rear of the robot.
8. Aggregating multiple sensors via sensor fusion.

The evolution of their UML USAR interface throughout the following years, was described in [17] and the most important milestones are visualized in *figure 4*. The first version (see topmost) is already based on the video-centric design approach, it includes a rear camera view and a map-based distance panel. The second version (see middle) reordered the panels and thereby gave the video stream more focus. In order to enhance awareness, they migrated the distance panel into a perspective view, in order to reduce cognitive load. The last improvement on the still joystick- and keyboard-based UML USAR interface (see bottom) includes the replacement of the distance panel through a zoom mode inspired panel. "Instead of boxes, lines are drawn around a scale model of the robot based on sensor information" [17]. These improvements were made to enhance the human-robot awareness. The next upgrade to the UML USAR interface involves the switch to a multi-touch based interface, being the main focus of Keyes' PhD thesis.

But before looking at this migration, we will have a quick look at RoBno's interface (see *figure 5*) and compare it with versions one to three of the UML USAR interface (see *figure 4*). Both interfaces are joystick-based and follow a video-centric approach. For RoBrno's interface they chose to use a fullscreen video mode, instead of working with multiple panels. This design decision was mainly due to the use of head-mounted displays instead of normal desktop monitors. Two similarities include the use of a cross-hair for signalizing the offset of the current view from the robot's driving direction and prioritizing and combining sensor information, being displayed on top of the video stream itself. Another interesting aspect is that they worked with two different display modes, one optimized for usability and one used for development purposes. The original interface, used for development purposes, turned out to be too complex and even being hindering for the four HRI guidelines introduced in *section 5*. The simplified version of the interface reduced the cognitive load and the amount of not-so-important data, which would normally overload the operator. The interface you see in *figure 5* represents the simplified version.

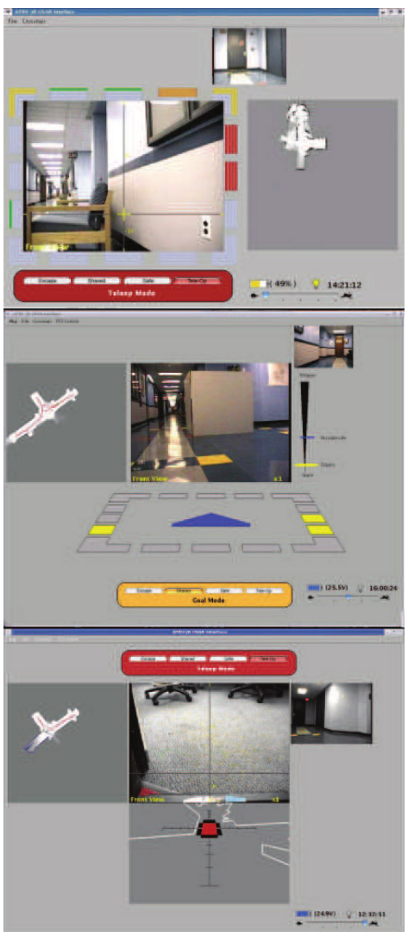


Fig. 4. Summary of the different joystick- and keyboard-based versions of the UML USAR interface [17].

Keyes et al. [17] describe the transition from the joystick- and keyboard-based UML USAR interface to an interface solely based on multi-touch technology (see *figure 4*). The goal of the further development was to build a visually identical interface to the previously joystick-based USAR interface. The main differences between both interface types are the ergonomics and the degree of direct manipulation. Due to the high degree of freedom of multi-touch interfaces the problem of "emergent behavior", describing "unintentional or surprising combinations of behaviors"[18], arises. This presents a higher demand on interface designers, since users with different backgrounds will try to use different ways of interaction (motions inspired from joystick, buttons, track-pads, piano keys, touch typing, and sliders). But besides this limitation the use of multi-touch interfaces brings significant improvements with it. Interactions are easier-to-learn and the ability of direct interactions are superior to other technologies (no additional layers of abstraction, resulting in a lower cognitive load). The results of Keyes' study is that touch-based interfaces do not degrade the overall performance, but actually slightly improve on it.

To complete this chapter, we will recap a few things to keep in mind. Lowering the cognitive load and enhancing awareness should be the most important aspects in developing a user-friendly interface. Optimizing the time needed for every single task, and providing help to the user, are all elements which increase the overall performance and usability of an interface. As seen in the previous examples, providing help in choosing the right autonomy mode, providing a simplified panel, using sensor fusion and displaying only relevant information directly near the viewer perspective, are all critical steps to improve the human-robot awareness and the humans overall awareness usability, depicted in *section 3*.

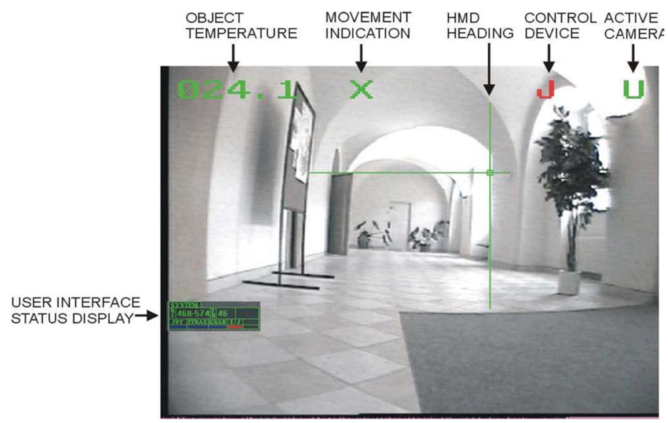


Fig. 5. RoBrno's interface, a strongly simplified interface, contradictory to the INEEL interface, presented in [4].

7 Trends

Before concluding this paper, we will have a brief look at some recent trends in HRI useful for designing interfaces. Throughout the last years a trend towards larger displays, richer video feeds (in terms of resolution and frame rate) and the use of multi-touch tabletops, tablets and smartphones as input devices, is observable. In addition to this trend, the recent explosion of touch interfaces and the herewith combined development of simpler and easier-to-use interfaces already began to show its influence on HRI, a development to which the iPhone has a significant share. Users desire simpler interfaces with more natural means of interaction, based on touch-, gesture- or voice-based technologies [18].

As already seen in Micire [18] the use of multi-touch interfaces adds a significant improvement to the user-friendliness and also slightly to the overall performance. Micire sees "a strong indication that this anatomy-based design is a good approach and additional improvements should be investigated". Throughout the last years the use of gestures and human voice commands, to interact with robots, has also increased. Stiefelhagen et al. [23] use speech recognition, multimodal dialogue processing, gesture recognition and the recognition of the users head orientation to create a whole new way of robots working alongside humans in the kitchen.

The work from Teller at al. [24] combines many of the present trends in a unique way. They developed a multi-ton robot forklift, which is able to operate alongside human personnel. The robotic forklift works with three types of inputs: a multimodal tablet, local sensing (containing distance sensors) and the direct interaction with humans through voice-commands and gestures. In their prototype, humans even have the ability to climb inside the robot during operation and overtake manual control. The combination of this variety of interaction types opens completely new possibilities. A more natural and intuitive collaboration between humans and robots is not far away anymore. When we look back to the three laws published by Iaac Asimov in 1941, that robots may not injure a human being, must obey their orders and protect their own existence, research got much closer to these in the beginning very visionary goals.

For further in depth information about current *state of the art* interfaces in HRI, have a look at the publication from Aryania et al. [2]. They deliver a good overview, discussing new trends in industrial robot controller user interfaces. Aroca et al. [1] work with smartphones to achieve smarter robots and a better usability. Cheng et al. [7] work on Kinect- and gesture-based interactions for HRI.

8 Conclusion

In this study we observed what has been accomplished in the area of human-robot interfaces so far, and evaluated what metrics are available to measure the performance of user interfaces interacting with robots. The goal was to summarize principles from literature, allowing the reader to build upon the knowledge of previous studies, with the ultimate goal of creating simpler and user-friendlier interfaces, leading to a higher performance for the end user. The focus of this paper was limited to acquiring situation awareness, yet many of the concepts and trends can be transferred to other human issues.

The significance of this paper is the recapitulation of several HRI interfaces from the last 15 years. At first we looked at what was a standard ten years ago, what problems they encountered, how the interfaces were iteratively improved and what is currently being referred to as state of the art. This knowledge is essential when wanting to build sleek and simple user interfaces for future robot applications. Without the knowledge of the past, the same mistakes are made over and over again, being the reason why we wanted to sum up the previous recommendations in literature. The goal was to find guidelines for the simplification of remote robot interfaces.

Furthermore we want to make an appeal for a stronger focus on the user-friendliness of robot interfaces. Not solely due to the increase in interaction performance, but due to recent developments of what is becoming standard for commercial users. There is a transition from the usage patterns of desktop computers, using a mouse and keyboard, towards touch- and gesture-based interactions. Users are already used to, and will soon be expecting, the standards of the web and smartphone industry. The trend is clear: Wherever more consumers are confronted with technological products (previously being reserved for technology-savvy people) the usability standard of the product rises. The prediction that HRI will become more mainstream and more consumers will be interacting with robots, clearly states the demand for the research and development of simpler ways to interact with remote robot interfaces. The current work of previously cited authors include experiments with direct manipulation (to be able to send a robot to a specific point) and with a big focus on multi-touch interaction methods [18]. My suggestions for the development of upcoming robots is to put a higher focus on the usability of the interfaces, since it will not only increase the overall performance of the robot, but also strengthen the customer loyalty and customer relationship.

References

[1] R. Aroca, L. Gonçalves, and P. Oliveira. Towards smarter robots with smartphones. *5th Workshop in Applied Robotics and Automation*, pages 1–6, 2012.

[2] A. Aryania, B. Daniel, T. Thomessen, and G. Sziebig. New trends in industrial robot controller user interfaces. pages 365–369, 2012.

[3] I. Asimov. I, robot. *a collection of short stories originally published between*, 1940.

[4] M. Baker, R. Casey, B. Keyes, and H. Yanco. Improved interfaces for human-robot interaction in urban search and rescue. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 3:2960–2965, 2004.

[5] D. J. Bruemmer, D. D. Dudenhoeffer, and J. L. Marble. Dynamic-Autonomy for Urban Search and Rescue. Technical report, AAAI, 2002.

[6] J. Chen, E. Haas, K. Pillalamarri, and C. Jacobson. Human-robot interface: Issues in operator performance, interface design, and technologies. (July), 2006.

[7] L. Cheng, Q. Sun, H. Su, Y. Cong, and S. Zhao. Design and Implementation of Human-Robot Interactive Demonstration System Based on Kinect. pages 971–975, 2012.

[8] J. L. Drury, J. Scholtz, and H. A. Yanco. Awareness in human-robot interactions. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 1, pages 912–918. IEEE, 2003.

[9] J. L. Drury and H. A. Yanco. Applying CSCW and HCI Techniques to Human-Robot Interaction. page 4, 2006.

[10] M. R. Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 32, pages 97–101. SAGE Publications, 1988.

[11] T. Fong, D. Kaber, J. Scholtz, and A. Schultz. Common Metrics for Human-Robot Interaction. 2004.

[12] T. Fong, C. Thorpe, and C. Baur. Multi-robot remote driving with collaborative control. *Industrial Electronics, IEEE Transactions on*, 50(4):699–704, 2003.

[13] M. Goodrich and D. Olsen. Seven principles of efficient human robot interaction. *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, 4:3942–3948.

[14] R. Hunter. How Robots Are Revolutionizing Our World. `http://www.popsci.com/trp-sponsored-article-analyst`, 2014. Accessed Mai 30, 2014.

[15] M. Jaschkowitz. Navigational Tasks in Human-Robot Interaction. *Media Informatics Advanced Seminar Secondary Tasks*, (LMU Munich), 2014.

[16] B. Kannan and L. E. Parker. Fault-tolerance based metrics for evaluating system performance in multi-robot teams. In *Proceedings of Performance Metrics for Intelligent Systems Workshop*, 2006.

[17] B. Keyes, M. Micire, J. L. Drury, and H. A. Yanco. *Improving Human-Robot Interaction through Interface Evolution*. Number February. InTech, 2010.

[18] M. J. Micire. *Multi-Touch Interaction For Robot Command And Control*. PhD thesis, University of Massachusetts Lowell, 2010.

[19] G. Mies. Military robots of the present and the future. *TECHNOLOGY*, 9(1):125–137, 2010.

[20] D. R. Olsen and M. A. Goodrich. Metrics for Evaluating Human-Robot Interactions. 2003.

[21] T. Simonite. Googles Strategiewechsel bei Roboterautos. `http://heise.de/-2213719`, 2014. Accessed June 25, 2014.

[22] S. M. Singer and D. L. Akin. A Survey of Quantitative Team Performance Metrics for Human-Robot Collaboration. (July):1–19, 2011.

[23] R. Stiefelhagen, C. Fogen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech, head pose and gestures. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3:2422–2427, 2004.

[24] S. Teller, M. R. Walter, M. Antone, A. Correa, R. Davis, L. Fletcher, E. Frazzoli, J. Glass, J. P. How, A. S. Huang, J. Sertac, K. Brandon, L. Nicholas, and R. Tara. A Voice-Commandable Robotic Forklift Working Alongside Humans in Minimally-Prepared Outdoor Environments The MIT Faculty has made this article openly available . Please share Citation Accessed Citable Link Detailed Terms A Voice-Commandable Robotic Forkli. 2014.

[25] C. Urmson. Just press go: designing a self driving vehicle. `http://googleblog.blogspot.de/2014/05/just-press-go-designing-self-driving.html`, 2014. Accessed June 20, 2014.

[26] J. Villbrandt. The Quadrotors Coming of Age. `http://illumin.usc.edu/162/the-quadrotors-coming-of-age/`, 2010. Accessed June 26, 2014.

[27] H. Yanco and J. Drury. "Where am i?" acquiring situation awareness using a remote robot platform. *IEEE International Conference on Systems, Man and Cybernetics*, 3:2835–2840, 2004.

Usability Aspects of Remote Camera Control in Film Production

Ozan Saltuk

Abstract— This paper focuses on the remote camera control issue in film production and tries to identify the usability problems for the camera operators who will control the camera-robots. We understand remote camera control as a subcategory of human-robot interaction. For the identification of the usability problems the general knowledge and recent findings in human-robot interaction field will be examined. Camera operators in film sets are the people who will use these robots, therefore information about film production and roles in the camera department will be introduced in order to understand the current working ways of camera operators. These two fields will be brought together in an effort to find out the potential and existing problems. The analysis will be done on two remote camera examples (IRIS and Scorpio) which have different working principles.

Index Terms—Control, Camera, Remote, Human-Robot Interaction, Remote Control, Teleoperation, Film Production

1 Introduction

When director Alfonso Cuarón spoke of his ideas for a new space film featuring zero gravity shots, he heard many discouraging words. Even fellow director David Fincher thought that his vision of weightlessness was not compatible with today's technology and advised him to wait five years [6].

However, the technology Cuarón needed already existed in the motion control robot called "IRIS". IRIS (seen in figure 1) is a new and high level tool for motion control by Bot&Dolly[1]. By using four IRIS robots, Cuarón was able to simulate zero gravity for his film and shoot the film the way he wanted. His film, entitled "Gravity", hit the theaters in fall 2013 and was a big success. It grossed 716 million dollars worldwide [2] and won academy awards for "Best Achievement in Cinematography", "Best Achievement in Directing" and five other categories [21].

Advances in robotics enable the building of a new generation of cameras and tools, which allow the filmmakers to film shots that were not possible before (for example the rolling camera motion used in Gravity). However, since this technology is new and high-level, it also introduces usability problems for the people who will have to operate them, for example in navigation related issues. Some remotely controlled cameras are navigated by operators via software. Camera operators, who are responsible for the actual usage of a camera in film sets, are not experts in working with such software or interfaces. Therefore, they can face usability problems while working with these robots.

Human-robot interaction is a fairly new and growing research field. Although robots are not as commonly used in daily life as science fiction writers imagined years ago, more and more research is being done in this field. Human-robot interaction has many application areas, ranging from urban search and rescue [39] or space exploration [13] to military [4].

By studying the production process in films and by applying the findings from human-robot interaction to remote camera control, the problems that are likely to arise in the future can be identified. This will help improve production and provide user support.

The aim of this paper is the identification of usability problems for the camera operators in film sets. Since these tools are robots, evaluation of the usability aspects require knowledge on human-robot interaction. Chapter 2 will be about the human-robot interaction field and contain information on important aspects in human-robot interaction. Chapter 3 will cover professional film production, the roles in film sets and will emphasize the similarities between the roles in Human-Robot Interaction and film production. Chapter 4 will bring these two subjects together, examine two different types of remote cameras and focus on the usability issues regarding robot-cameras. A discussion about the topic will be given in chapter 5. The last chapter will summarize the paper.

- *Ozan Saltuk is studying Media Informatics at the University of Munich, Germany, E-mail: ozan.saltuk@campus.lmu.de*

- *This research paper was written for the Media Informatics Advanced Seminar 'Special Aspects of Usability', 2014*

[1]http://www.botndolly.com/system

Fig. 1. IRIS by Bot & Dolly, the motion-control robot used in Gravity[2]

2 Human-Robot Interaction

In this chapter general knowledge for describing and evaluating human-robot interaction will be given so that it can be applied to the field of remote camera control (see chapter 4). This general knowledge will include the levels of automation, programmable robots, the task metrics for robots, the roles in human-robot interaction and the operator performance issue.

2.1 Levels of Automation in HRI

Parasuraman et al. described automation in a general context as "to what degree the robot can act on its own accord" and defined a scale consisting of ten different levels of automation [27]. Their scale ranged from the human controlling the robot completely (teleoperation) to the robot being fully autonomous.

Goodrich and Schulz argued that when creating a scale of autonomy, the level of interaction between the human, the robot and the degree to which each is capable of autonomy is also important. They defined a new scale with an emphasis on mixed-initiative interaction, which describes "a flexible interaction strategy, where each agent can contribute to the task what it does best" [1, 14].

Their scale, which can be seen in figure 2, was optimized for human-robot interaction. This scale ranges from direct control to dynamic autonomy.

[2]http://unifiedpoptheory.com/wp-content/uploads/2012/06/Bot_Dolly_IRIS-e1341358285329.jpg

Automation ranges from teleoperation (the robot offers no assistance, the operator does it all) to full autonomy (the robot decides everything and acts autonomously, ignoring the human).

The level of required autonomy depends on the task the robot was designed to perform. Planetary surface exploration robots like the Mars Curiosity should be able to decide on certain actions themselves [35], since the territory they operate in is not fully known and real time communication with a human operator is not always possible. However, robots used in other fields, for example camera-robots in film sets, do not necessarily need any such decision capabilities, since they operate in a known environment and they have to follow orders given by human operators without any change.

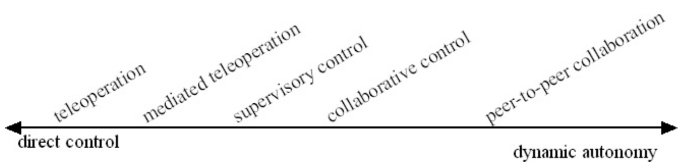


Fig. 2. Levels of autonomy with emphasis by Goodrich and Schulz [14].

2.2 Programmable Robots

The levels of automation that were introduced in the previous chapter were for non-industrial robots. Another class of robots are the industrial robots, which are designed to assist humans in industrial settings [14]. These types of robots are programmed by the humans beforehand to reproduce the programmed movement once or more depending on the task.

The programming methods of industrial robots are divided into two categories: online and offline programming [26]. In online programming, the robot is manipulated by the human to the desired positions and orientations. It is suitable for uncomplicated tasks with simple geometry but the performance or precision is dependent of the motor skills of the operator. In offline programming, the programming is done on a computer via software, enabling more precise movements. It also has other advantages in terms of improved performance and reduced downtime [5], but requires experience with the software or tools.

2.3 Task metrics in Human-Robot Interaction

In order to analyze and evaluate human-robot interaction, Steinfeld et al. listed five tasks as metrics for task-oriented mobile robots [34]. They chose these tasks because they can be applied on robots with different levels of autonomy. These categories are *perception*, *navigation*, *management*, *manipulation* and *social*. Navigation means the movement of the robot, perception is understanding the environment, management describes the coordination of actions, manipulation stands for the robots interaction with the environment and social measures social interactions.

2.3.1 Perception

Steinfeld et al. describe perception as "the process of making interferences about distal stimuli (objects in the environment) based on proximal stimuli (energy detected by sensors)" [34]. They define two main tasks involved in perception: interpreting sensed data and seeking new sensor data. They divide perception metrics in two categories. Passive perception means the interpretation of the sensor data, whereas active perception measures performance on recognition tasks.

2.3.2 Navigation

In most human-robot interaction scenarios featuring a mobile robot, navigation is a fundamental issue. Steinfeld et al. divide navigational tasks into three categories, which are global navigation, local navigation and obstacle encounter [34]. Global navigation means that the overall location and the environment of the robot should be understood. A robot designed to operate in deserts might not have the necessary tools or parts to operate in other environments, for example urban terrain. Local navigation is the second category and describes a more detailed information about the environment for example the stairs or people. The last category is obstacle encounter. Obstacle encounter describes how the robot should handle situations where an obstacle prevents it from navigating as planned.

2.3.3 Management

Management means the coordination of actions of humans, robots and the used resources. Using the most suited robot for a task within an optimal time frame would be an example of good management.

2.3.4 Manipulation

Robots working in unknown or unmapped environments may need to interact with the environment. These interactions can be motions such as grasping, pushing, lifting etc.

2.3.5 Social

Robots designed for social tasks need to interact with humans. Such robots have a wide potential range of application fields from entertainment to health care [16].

2.4 Roles in Human-Robot Interaction

The roles in human-robot interaction describe what kind of a relationship the human has with the robot. Scholtz originally defined three roles: supervisor, operator and peer. She then added a mechanic role and divided the peer role into a bystander and teammate role [29, 31].

2.4.1 The Supervisor Role

The supervisor role in human-robot interaction is quite similar to the supervisor role in human-human interaction, whereas the supervisor monitors the situation but does not need to operate the robot by him/herself. The supervisor's main job is to evaluate the situation and check if the tasks are being completed efficiently and correctly. In cases where there are more than one operating robot, the supervisor may need to work with multiple robots at the same time.

2.4.2 The Operator Role

The operator controls the robot. For any type of robot which has the least possible amount of automation (teleoperation), the operator role is essential, whereas robots with full autonomy do not require operators. The operator needs to have at least basic knowledge about the robot's capabilities and the interaction interface. Lack of such knowledge may result in unsatisfactory results or extended task completion times.

2.4.3 The Mechanic Role

The mechanic makes physical changes or alterations on the robot, such as repairs or adjustments. Therefore he/she should be located in the same environment as the robot.

2.4.4 The Bystander Role

The bystander does not always need to interact with the robot, but he/she needs to be able to predict the robot's behavior so that the robot can achieve its task without any interruptions.

2.4.5 The Teammate Role

The teammate interacts with the robot in order to achieve a certain goal together. Scholtz explains in [29] that the term teammate should not imply that the robot and the human is equal as the name might suggest, rather both the human and the robot contribute skills according to their ability.

2.5 Operator Performance

In this section the parameters for operator performance measurement according to Steinfeld et al. will be presented [34].

2.5.1 Situational Awareness

First issue in operator performance is situational awareness. Endsley describes situational awareness simply as "knowing what is going on around you" [8]. Basically everyone can have a degree of situational awareness. As Endsley suggests, someone sitting idly under a tree can have an idea of what is going on around him/her, but the focus should be set on people who need a certain degree of situational awareness while trying to achieve a goal or perform a task. The kind of information the subject needs is also relevant. For example for a firefighter entering a burning building it is important to know how many people might be there and where they are likely to be, but it is irrelevant to know who the building is owned by. By taking account this extended information, Endsley makes a more formal definition of situational awareness as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [7]. Scholtz argues that although performance and situational awareness are related, they are not directly correlated [29] (Meaning it is possible for someone to have good situational awareness and perform poorly or vice versa) .

Situational awareness consists of three levels [9]. Level one describes the perception of cues (simple recognition of items). Level two is the ability to bring the pieces of information together and evaluate relevance of the information to the task. Level three is being able to predict future events based on the perception and evaluation of the current situation.

2.5.2 Mental Model

Another important issue is the mental model. Klimoski and Mohammed refer to the term mental model as "a general class of cognitive constructs that have been invoked to explain how knowledge and information are represented in the mind" [18]. In simpler words, mental models describe how a person represents a situation in his/her mind. Mental model accuracy is important since a correct mental model can reduce cognitive load and enhance learning [19].

2.5.3 Workload

The last issue in operator performance is (mental) workload. Based on Moray's mental workload concept [22], Wickens describes mental workload as "the relation between the (quantitative) demand for resources imposed by a task and the ability to supply those resource by the operator" [36]. In simpler words, workload can be understood as the amount of work the person has to deal with. There are a number of factors influencing workload. Automation of the robot is one of them. The more the robot is able to decide for itself, the less work the operator has to do. Whether the robot decides well is another issue. Also the human-robot ratio in a team is an important workload parameter. In application scenarios where a single operator controls multiple robots, the workload is significantly higher. Prewett el al. described guidelines and propositions for reducing operator workload in human-robot interaction [28].

The most well known tool for measuring workload is the NASA Task Load Index (NASA-TLX), which uses the following six dimensions: mental demand, physical demand, temporal demand, performance, effort and frustration level [16].

Rosemarie E. Yagoda developed a tool for measuring workload specifically in human-robot interaction [38]. Her argument was that the conventional measurement tools like the NASA-TLX did not assess the workload attributed to the system configuration and did not consider the robot-team dynamics. She used five attributes for the evaluation of workload: task, system, team process, team configuration and context. It should be noted that this tool was designed for the operators to evaluate their own workload.

3 Film Production

In 2013, worldwide box-office grossed more than 30 billion dollars [10]. A research made by Stephen Follows based on the data from the international movie database (IMDb) shows that in the 50 highest grossing films for each year in the past 20 years (from 1994 to 2013 with a total of 1000 films), there was an average of 588 crew working in a film [12].

3.1 Roles in the Camera Department

Follows' research also shows that in average 55 of the 588 people in the crew work in the camera and electrical department (IMDb lists these two departments together), making the camera and electrical department the third biggest department (sharing the third place with the art department) in a film set. The biggest is visual effects with 156, the second "other crew" with 103. 2013 film "Now You See Me" holds the record for the largest camera and electrical department with 334, six times the average number.

In this chapter the roles in the camera department will be examined in order to bring together the roles in human-robot interaction as described in the previous chapter and the roles that will be introduced here. In this way the knowledge from the human-robot interaction research field can be applied to the relevant application area. The major roles are as follows [11]:

3.1.1 Director of Photography/Cinematographer

The Director of Photography is in charge of the camera and lightning departments. The major decisions on lightning and framing are made by him/her in collaboration with the film's director.

3.1.2 Camera Operator

The Camera Operator receives instructions from the director of photography or the director and operates the camera accordingly. Traditionally one camera operator does not operate more than one camera at the same time.

3.1.3 First Assistant Camera

The First Assistant Camera's (also called Focus Puller) job is to make sure the camera is in focus while shooting. Focus change is needed if the actors or the objects that are being filmed move towards or further away from the camera.

3.1.4 Second Assistant Camera

The Second Assistant Camera (Clapper Loader) uses the clapperboard before acting starts in each take. He/she also loads the blank film into the camera before the takes if they are working with an analog camera. He/she is also responsible for the logs kept for the film.

There is also a number of minor roles in the camera department. Loader (takes the film from light-tight canisters and puts them into camera magazines), Camera Production Assistant (basically a trainee), Digital Imaging Technician (responsible from the digital camera), Data Wrangler (transfers the data from the camera to the computer), Steadicam Operator, Motion Control Operator (operates a motion control rig which repeats camera moves for special effects use) and Video Split/Assistant Operator (responsible for the system the director uses to watch videos after each take).

3.2 Similarities of Roles in HRI and Film Production

In order to evaluate remote cameras in film sets in a human robot interaction context, we map the roles Scholtz defined in [31] to the roles in film sets.

The role with the most authority in human-robot interaction is the supervisor role. The supervisor's correspondent in a camera department would be the director of photography/cinematographer. They are both responsible for monitoring the situation and making sure everything goes as planned. They do not operate the device physically, instead they have a more thinking and planning role. One could argue that the actual director of the film should be mapped with the supervisor role. This would also be reasonable, since the director is in a more advanced position in the hierarchy than the director of photography. However, since we are just talking about the camera department in our task, the director is not really fitting, because he/she is the head of all creative departments.

The second role, the operator, would be mapped to the camera operator (or in cases where there is a motion control operator to the motion

Fig. 3. Scorpio Camera Crane[3]

control operator) role as the name suggests. The focus of this paper will be on the camera operator and the operator roles. They are both responsible for the physical usage of the robot, and the usability problems concern them directly. The major difference is that the operator in human-robot interaction operates the robot remotely, whereas traditionally the camera is held and/or moved by the camera operator.

The role of the mechanic can be mapped to the roles of the camera assistants. Both assistants do not necessarily have the skills required to fix the camera when it is damaged, but they both make physical alterations on the device. The first assistant camera changes the lens focus manually, whereas the second assistant camera loads the blank film into the camera before takes. Since without these roles the camera would not function as wished, it is suitable to map the role of the mechanic to these two roles.

For the teammate role, the description should be remembered. The teammate is described as a person who interacts with the robot in order to achieve a certain goal together. This could imply that the rest of the camera department would be counted as teammates to the robot, since their goal is the same (filming the shot) and they assist each other according to their abilities. However, during filming there is no real interaction between those roles and the camera-robot while filming that clearly fits in this category.

4 Remote Camera Control

In this chapter, knowledge from chapters 2 and 3 will be applied to the remote camera control field in order to identify potential usability issues. Two camera-robots with different usage methods will be examined in detail. One of them, IRIS, is preprogrammed and is closer to the industrial robots in human-robot interaction sense while the other one, Scorpio, is teleoperated [33]. We will assume that in both scenarios one camera operator will not control more than one camera simultaneously, as in traditional film making.

4.1 The Operator Role and Usability

The operator is the most relevant role when identifying problems, since he/she is the person who will do the actual remote camera control. Issues which were introduced generally in the previous chapters will be addressed more specifically in the following.

The United States Bureau of Statistics suggest that many camera operator positions require a bachelor's degree in film related subjects [3]. Since film making shifted from analog to digital in the last few decades, the working principles of cameras also changed. Human skills as communication, creativity or hand-eye coordination are still valid requirements for a good camera operator, but the operator should also have an understanding of how the digital cameras work.

[3]http://www.southeastcameracranes.com/gallery/albums/gallery/img_8870.jpg

IRIS adopts the offline programming method which was introduced in 2.2 and uses Maya as programming platform. Bill Galusha, a senior producer from Bot&Dolly said in an interview that anyone who is a character animator can use IRIS [15]. However, Scholtz mentions that "the operator must be a skilled user, having knowledge of the robotic architecture and programming" [29]. It could be expected that film school students learn about new technologies and software regarding film. Unfortunately, this only applies to camera operators who went to film school recently. As in most fields, finding a good paying job as a camera operator usually requires good experience. Therefore not all professional camera operators would be familiar with using software, let alone have knowledge of the robotic architecture and programming.

Another camera-robot which is more common and closer to the traditional camera usage in technological spectrum is the Scorpio Telescopic Camera Crane (seen in figure 3). Scorpio is a system developed by Southeast Camera Cranes. Unlike IRIS, Scorpio uses real time teleoperation as the movement of the camera is not preprogrammed. The camera is positioned on a movable crane and the operator stands in front of a control panel. On the control panel there is a monitor where the operator can look through the camera's view and two handles which are used to move the crane or change the angle of the camera.

Of the five metrics introduced in 2.3, two are applicable to the usability issues in operator level remote camera control field: perception and navigation. Management is not a usability issue in operator level, the allocation of the right robot is not the operator's responsibility. Manipulation also is not applicable, because in film sets everything is planned with the camera and camera route in plan. The sets are carefully designed by set designers. Therefore the camera-robot shouldn't have to make any changes in the environment. Social interaction is also an unrelated issue in this field, since the remote cameras are not "social robots".

4.2 Perception

The emphasis will be on perception related issues. Situational awareness, mental model and workload related issues in this particular context will be discussed here.

4.2.1 Situational Awareness

For the operator role, Scholtz hypothesized that the operator role in human-robot interactions need to have the following information for good situational awareness [29]:

1. The robot's world model
2. The robot's plans
3. The current status of any robotic sensors
4. Other interactions currently occurring
5. Any other jobs that are currently vying for the operator's attention (assuming it is possible to service more than one robot)
6. The effects of any adjustments on plans and other interactions
7. Mission overview and any timing constraints

Most of these points are applicable to remote camera control, with the exception of points 2, 4 and 5. The robot's plans (2) is not relevant for IRIS, since the robot repeats the movement which was programmed by the human and also not relevant for Scorpio, since it is teleoperated. In both cameras the human makes the plans for the robot and the robot's plans are exactly the same as given. Also the robot will not have any other interactions currently occurring, rendering the fourth point irrelevant. The fifth point is also not applicable, since the assumption made is not true in this case. Therefore, reaching good situational awareness is easier in remote camera control in comparison to other fields of human-robot interaction.

As Scholtz mentions, Murphy and Rogers defined three drawbacks to telesystems [23], which are:

1. The need for a high communication bandwidth for operator perception and intervention

2. Cognitive fatigue due to repetitive tasks
3. Too much data and too many simultaneous activities

Operators of IRIS and Scorpio have different perception issues. IRIS is preprogrammed, which means that the operator does not have the opportunity to make live alterations but which also means that everything should be modeled beforehand.

Operators of Scorpio work in real-time and see the view of the camera through a monitor. The latency time is the biggest issue in this case, since the operator make changes in real time. This is exactly the same as the first drawback Murphy and Rogers defined. The second point is also relevant, since the operator may need to repeat the same motion over and over again. These problems are not relevant in case of IRIS. The third point is dependent on the situation in the film set.

With Scorpio, the operators also have a reduced view of the set if they are not co-located with the robot. Although not likely, it can be that they should teleoperate the device from another location, in which case the situational awareness would be low, since they would be limited to the camera's view. These kind of problems are more relevant in other fields of human-robot interaction, for example urban search and rescue [30].

4.2.2 Mental Model

A number of parameters can influence human performance. In the case of IRIS, it may be more difficult to build a correct mental model of the device operation. Designing 3D routes in 2D interfaces is challenging even if the person has some experience with it. Computer screens are two dimensional interfaces, designing a route can be most easily done in a bird's-eye view, which means that the route would always be in the same height. In the case of Scorpio it is much simpler, because learning by doing is supported and it works like a traditional camera, whereas in IRIS it takes more time to program the route, upload it to the robot and see if it behaves as planned.

4.2.3 Workload

It would be possible to apply Yagoda's workload measurement tool which was introduced in 2.5.3 for human-robot interaction in a hypothetical situation. The task is to control the camera correctly and film the wished shot. The system consists of the robot and the interface. As mentioned in mental model, the operator has to have a good understanding of the system. The team process would be the operator working successfully with the robot. As stated before, it is assumed that one operator controls a single robot, therefore the team configuration would be one. The context would be the film shooting scenario.

The measurement of each attribute is highly dependent on the operator's experience and skills. For a camera operator with no experience with remote controls or interfaces, the amount of workload for attributes such as task, system and team process are extremely high. IRIS would be a bigger challenge for a camera operator since the movement of the camera is programmed beforehand. It would also be a bigger challenge for the actors. In the example of Gravity, all of the movements of the actors were predefined, which meant they needed precise choreography and did not have much room for improvisation [37].

In comparison to other human-robot application fields, the team configuration is an issue that would reduce the workload of the operator. The context attribute is also low, since there would be no or minimal external factors disturbing the operator.

4.3 Navigation

Applying Steinfeld's navigation model from 2.3.2 to remote camera control shows interesting results. The first point of the model, global navigation, is vital for remote cameras in film sets. Local navigation is also important. For IRIS, the camera-robot should follow the orders that are given to it exactly, which means that the set should be modeled with high detail. For Scorpio, local navigation is more problematic for the operator if the operator is not in the same location as the camera.

In other application fields of human-robot interaction, the third category of Steinfeld's model (obstacle encounter) is a major issue. For example in planetary surface exploration the robot has to deal with obstacles frequently, since the terrain it operates on is not fully known. Also in application fields like urban search and rescue, the robot often needs to navigate in or around obstacles such as collapsed buildings, which cannot be mapped beforehand. However, in remote camera control such problems are not relevant. When designing a film set, the position of the camera is one of the most important issues. The route of the mobile camera is also considered before shooting, making the obstacle problem irrelevant. Robot-cameras have a small range of automation. This means that they can not change their course by themselves. If they could, it would change the whole shot, resulting in unwanted or unusable shots.

4.4 Remote Cameras as Products of Design Science Research

Remote Camera Control technology can be seen as a product of design and creation research or design science paradigm. The design-science paradigm derives from engineering [32] and is a problem solving algorithm. It seeks to "create things that serve human purposes" [20]. These products, also called artifacts, are created in order to solve a problem that exists in an environment using the foundations or methodologies that exist in the knowledge base. These devices are built and evaluated until a satisfactory product is created, as Hevner et al. described in their research framework in [17] .

Defining remote camera control devices as a product of design science, sheds lights on what kinds of problems may arise in this area. Oates lists a disadvantage for design and creation research evaluation that is highly relevant to our issue [25]. She argues that the success of an IT artifact is dependent on the researchers being present and the absence of the them may result in the system not working efficiently. This could be highly impractical; imagine in order to do video chat with your smartphone, you need one of developer's help.

One of the biggest challenges in this issue is that the technology being introduced is quite different from the technology it will succeed. Although both are used for the same purpose (filming) the ways they are used differ. The new technology is used remotely via software or with handles. The researchers of the technology being absent may result in an inefficient usage of the camera by creating major workload.

5 Discussion

Film production is a gigantic industry and innovative filmmakers are always trying to use new technologies that could draw potential viewers to theaters. While some of these innovations, for example sound in films, created new jobs in film production, others, for example the invention of color film, changed the way how people do their job.

In chapter 4, we have seen what might cause problems for camera operators working with remote cameras. I think the biggest and most important issue is lack of experience in working with computer interfaces. As seen in 4.2.3, when the operators are not familiar with such interfaces, the workload is extremely high, slowing down the production process. This is expected, since camera operator is traditionally a "hand-to-hand" job, whereas remote camera control is not. It is also not easy to practice with these robots, since they are extremely expensive and a normal camera operator would not have easy access to one.

Technology is advancing really fast, and remotely controlled cameras might replace normal cameras in film sets. Maybe in 15 years, maybe more, but if they do, the requirements for camera operator jobs will also change. This will lead to new courses in film schools and the experienced camera operators will have to learn to use such technologies elsewhere. Film studios opening courses or workshops for the camera operators could be one possibility in achieving this.

The more advanced remote cameras will become, the cheaper the not-so-advanced versions will be (as it is with every technological device in the market). This means access to them will also be easier. Just a few decades ago, having a portable camera was a big luxury, but then in the '90s handheld portable cameras got really popular and easy to use. Today most people basically carry around a high definition camera with them (smartphones). It is not unreasonable to think that remotely controlled cameras will also be affordable.

The products may change. The examples that were introduced in this paper are the pioneers in the field. They do not make decisions for themselves, but it is not unreasonable to think of a camera who navigates the set by itself and finds the optimal shot under given parameters and shoots it. They might even react to the actors. For example if the actor decides to improvise and do something unexpected, the camera can find the optimal shot for the situation and shoot the scene. Remote cameras can also be "social robots" in the future. It may be possible to give orders to remote cameras and the cameras could answer back. However, even if the products change, the basic principles will be the same.

One could argue that although film production is such a big industry, the pressure on the camera operator is not as high as it is in some other fields of human-robot interaction. This does not mean that the camera operators have an easy or stress-free job but the camera operator controlling the robot with less than 100% efficiency would lead to delays in production, whereas in fields like urban search and rescue or health care, inefficient usage could lead to fatal results. The goals in remote camera control can be quite different. For example the Mars Curiosity, a 2.6 billion dollar space exploration robot, has 17 cameras [24]. It is designed for scientific goals whereas in film sets the goals are aesthetic. However, every aspect of each remote camera control should be studied carefully so that it can be applied to other fields if needed.

Since remote cameras are fairly new and rather a commercial product, there is not much scientific research done on them. They are also extremely expensive, therefore it is not always possible to test one. Without having access to the product and with such a lack of material, the research on such devices is also extremely hard. Gravity was a pioneering film in remote camera control and there is not much data on the operator's experience or problems which can be used as a starting point.

It is not possible to apply all principles from general human-robot interaction to our problem. Most frameworks use metrics such as the level of decision capabilities of robots or the communication with the human, but in film production with this stage of technological advancement, these topics are only partly relevant. This means most of the research from human-robot interaction is not suited and special research for this context is needed.

6 Conclusion

Advances in technology change a lot of aspects in professional and daily life. The conventional ways of doing specific jobs might change drastically with the introduction of robots. Although we are not yet used to having robots in our daily lives, more and more robots take part in professional fields. This does not necessarily mean that the robots are going to take over peoples jobs, but people in some fields should learn to work with or control robots.

One example for such a field is film production. Camera operators in film sets do not traditionally use computer interfaces or software to control the camera. This produces special usability problems for them. Remote cameras which are controlled by software are not easy to use by camera operators in film sets since the lack of experience creates a high workload for them. Learning by doing is not always an easy solution, since the devices are ofttimes extremely expensive.

In this paper the remote camera control issue in film production was presented based on two different types of cameras. IRIS, which is preprogrammed and used via software, creates a high workload for the traditional camera operators because of the lack of experience. Scorpio, which is teleoperated, creates other usability problems such as latency time or reduced situational awareness when the camera and the operator are not co-located. However, some major issues with preprogrammed robots such as lack of experience is not highly problematic, since learning by doing is supported.

References

[1] J. Allen, C. Guinn, and E. Horvtz. Mixed-initiative interaction. *Intelligent Systems and their Applications, IEEE*, 14(5):14–23, 1999.

[2] B. Barnes. Author of the book gravity sues warner bros over credit. `http://www.nytimes.com/2014/04/30/business/media/author-of-the-book-gravity-sues-warner-bros-over-credit.html/`, April 2014. [Online; accessed 27-May-2014].

[3] U. D. o. L. Bureau of Labor Statistics. How to become a film and video editor or camera operator. `http://www.bls.gov/ooh/media-and-communication/film-and-video-editors-and-camera-operators.html/`, January 2014. [Online; accessed 27-May-2014].

[4] J. Y. Chen, E. C. Haas, K. Pillalamarri, and C. N. Jacobson. Human-robot interface: Issues in operator performance, interface design, and technologies. Technical report, DTIC Document, 2006.

[5] F. S. Cheng. *Advanced Techniques of Industrial Robot Programming, Advances in Robot Manipulators*. INTECH, 2010.

[6] S. C. David and D. McNary. Alfonso cuaron returns to the bigscreen after seven years with gravity. `http://variety.com/2013/film/news/alfonso-cuaron-returns-to-the-bigscreen-after-seven-years-with-gravity-1200596518/`, September 2013. [Online; accessed 27-May-2014].

[7] M. R. Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 32, pages 97–101. SAGE Publications, 1988.

[8] M. R. Endsley. Direct measurement of situation awareness: Validity and use of sagat. *Situation awareness analysis and measurement*, 10, 2000.

[9] M. R. Endsley. Theoretical underpinnings of situation awareness: a critical review. In M. R. Endsley and D. J. Garland, editors, *Situation Awareness Analysis and Measurement*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2000.

[10] FilmL.A.-Research. 2013 feature film production report. http://www.hollywoodreporter.com/sites/default/files/custom/Embeds/, 2014. [Online; accessed 25-May-2014].

[11] FilmVictoria. Film & tv production roles and departments. `http://www.film.vic.gov.au/__data/assets/pdf_file/0010/4897/Crew-Roles-and-Departments-Guide.pdf`. [Online; accessed 28-May-2014].

[12] S. Follows. How many people work on a hollywood film? `http://stephenfollows.com/how-many-people-work-on-a-hollywood-film/`, February 2014. [Online; accessed 25-May-2014].

[13] T. W. Fong and I. Nourbakhsh. Peer-to-peer human-robot interaction for space exploration. In *AAAI Fall Symposium: The Intersection of Cognitive Science and Robotics: From interfaces to Intelligence*. AAAI, October 2004.

[14] M. A. Goodrich and A. C. Schultz. Human-robot interaction: A survey. *Found. Trends Hum.-Comput. Interact.*, 1(3):203–275, Jan. 2007.

[15] R. Hamden. Bot & dolly - iris cinema robot - cinegear 2012 (digital cinema training). `https://www.youtube.com/watch?v=kaX1S_dqVX4`, June 2012. [Online; accessed 28-May-2014].

[16] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[17] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Q.*, 28(1):75–105, Mar. 2004.

[18] R. Klimoski and S. Mohammed. Team mental model: Construct or metaphor. *Journal of Management*, 20:403–437+, 1994.

[19] J. A. Macedo, D. B. Kaber, M. R. Endsley, P. Powanusorn, and S. Myung. The effect of automated compensation for incongruent axes on teleoperator performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(4):541–553, 1998.

[20] S. T. March and G. F. Smith. Design and natural science research on information technology. *Decis. Support Syst.*, 15(4):251–266, Dec. 1995.

[21] R. Miller. 'gravity' wins seven oscars, including best director. `http://www.theverge.com/2014/3/2/5464334/gravity-2014-oscar-wins`, March 2014. [Online; accessed 27-May-2014].

[22] N. Moray. *Mental Workload: Its Theory and Measurement*. Plenum Press, New York, 1979.

[23] R. R. Murphy and E. Rogers. Cooperative assistance for remote robot supervision. *Presence*, 5(2):224–240, 1996.

[24] NASA. Eyes and other senses. http://mars.jpl.nasa.gov/msl/mission/rover/eyesandother/. [Online; accessed 28-May-2014].

[25] B. J. Oates. *Researching Information Systems and Computing*. Sage Publications Ltd., 2006.

[26] Z. Pan, J. Polden, N. Larkin, S. Van Duin, and J. Norrish. Recent progress on programming methods for industrial robots. *Robotics and Computer-Integrated Manufacturing*, 28(2):87–94, 2012.

[27] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. A model for types and levels of human interaction with automation. *Trans. Sys. Man Cyber. Part A*, 30(3):286–297, May 2000.

[28] M. S. Prewett, R. C. Johnson, K. N. Saboe, L. R. Elliott, and M. D. Coovert. Managing workload in human-robot interaction: A review of empirical studies. *Comput. Hum. Behav.*, 26(5):840–856, Sept. 2010.

[29] J. Scholtz. Theory and evaluation of human robot interactions. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 5 - Volume 5*, HICSS '03, pages 125.1–, Washington, DC, USA, 2003. IEEE Computer Society.

[30] J. Scholtz, J. Young, J. L. Drury, and H. A. Yanco. Evaluation of human-robot interaction awareness in search and rescue. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, volume 3, pages 2327–2332. IEEE, 2004.

[31] J. C. Scholtz. Human-robot interactions: Creating synergistic cyber forces. In *In Proceedings from the 2002 NRL Workshop on Multi-Robot Systems*, pages 177–184. Kluwer Academic Publishers, 2002.

[32] H. A. Simon. *The Sciences of the Artificial (3rd Ed.)*. MIT Press, Cambridge, MA, USA, 1996.

[33] SoutheastCameraCranes. Scorpio telescopic camera crane: Pan bars & mini head). http://www.southeastcameracranes.com/gallery/videos/panbars.txt.php, June 2011. [Online; accessed 20-June-2014].

[34] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI '06, pages 33–40, New York, NY, USA, 2006. ACM.

[35] G. Webster. Nasa's mars curiosity debuts autonomous navigation. http://www.jpl.nasa.gov/news/news.php?release=2013-259, August 2013. [Online; accessed 28-May-2014].

[36] C. D. Wickens. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2):159–177, 2002.

[37] F. Wickman. How they made gravity. http://www.slate.com/blogs/browbeat/2014/01/03/how_they_made_gravity_behind_the_scenes_video_shows_the_pioneering_light.html, January 2014. [Online; accessed 27-May-2014].

[38] R. E. Yagoda. Development of the human robot interaction workload measurement tool (hri-wm). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 304–308. SAGE Publications, 2010.

[39] H. A. Yanco, M. Baker, R. Casey, B. Keyes, P. Thoren, J. L. Drury, D. Few, C. Nielsen, and D. Bruemmer. Analysis of human-robot interaction for urban search and rescue.

Do not just use it, feel it: Experience Design - Beyond Usability

Markus Wirth

Abstract— The aim of this paper is to define the terms usability and experience design and to work out their features. Regarding usability a detailed explanation of the aspects efficiency, effectiveness, satisfaction and context of use is given. Further the evaluation-method MUSiC is introduced. The second part explains the term experience Design. In this context the term experience is explained and two approaches of user experience are introduced and examined. The first theory is about hedonic and pragmatic quality suggested by Mark Hassenzahl and the second one the three levels of processing from Don Norman. Another focal point will be the relationship between usability and user experience. In this area three approaches are introduced. Depending on the point of view usability is either a part of user experience or the other way around. The third theory which is presented in this context supports the perspective of seeing user experience as an extended term of satisfaction. Furthermore two examples show that it can be useful to neglect usability aspects to create and gain a particular experience.

Index Terms—User Experience, Experience Design, Usability

1 Introduction

By now technology is omnipresent in your everyday life and became indispensable. Due to this fact, the research field Human-Computer-Interaction is getting more and more attractive to people. By developing devices to support our daily life, it is essential to consider different aspects like context of use, efficiency and effectiveness. Hence, a product has to be useful. This usefulness can be guaranteed by providing utility on the one and usability on the other hand. With the help of measuring utility, the sufficiency of functional aspects for a device can be examined. It gives the answer to the question ”What can I do with this product?”. On the contrary usability is focusing on how a product can be used. Is the use of it efficient and effective, users will develop a positive attitude towards the product, this is, they will be satisfied [7].

In his book ”Emotional Design” Don Norman [35] talks about the collection of teapots he owns. One of them is completely unusable, because the handle is on the same side as the spout. The second one is due to his unique squat and chubby nature appealing. The third teapot is very functional and therefore difficult to use. Even though not using them on a daily basis, he is very attracted to his teapots. Norman loves watching them and comparing their contrasting shapes. Further they all tell a story out of his life. While one is reflecting his past, his crusade against unusable objects, the other reflected his future, his campaign for beauty. The third one represents a fascinating mixture of the functional and the charming [35]. This example of Don Normans teapots shows, that there is something more than functionality which has to be considered by designing a product. Although at least one of his pots is not usable, he nevertheless is attracted to it and likes to own it and sometimes presents the teapot to friends. This is due to the fact that emotions and experiences are connected to these products. To understand why in nowaday’s product development besides usability the term experience including user experience also plays a decisive part, these aspects must be examined in greater detail.

This paper marks out the term usability. Therefore the term usability is discussed and defined. Additionally is shown how usability can be measured and the correlation between aspects of measuring are pointed out. Finally the evaluation-method ”MUSiC” is exhibited. Beyond usability, experience design is discussed in the second part of this elaboration. User experience is defined and the properties of experience are pointed out. Finally the relation between usability and user experience is clarified.

• *Markus Wirth is studying Human-Computer-Interaction at the University of Munich, Germany, E-mail: markus.wirth@campus.lmu.de*

• *This research paper was written for the Media Informatics Advanced Seminar ’Special aspects of usability’, 2014*

2 Usability

Whereas before the 1980s the use of computers was reserved to a majority of special trained professionals, during this time the importance of personal computers grew. Due to the difficulty of use most people associated the interaction with a computer with frustration and anxiety. Quickly it became clear that this lack of usability has to be changed [12]. Therefore the first concepts for usability have been developed by B. Schackel [40] and formalized into a definition from J. Bennett [3, 41]. Today there are at least three definitions. The Usability Professionals’ Association recommended to view usability as

> ”The degree to which something - software, hardware or anything else - is easy to use and a good fit for the people who use it.” [1]

As written in ISO 9241-11 the ISO software ergonomics committee defined the term usability as

> ”The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” [7]

Thus the main goal of usability is to design effective and efficient products.

Whereas this definition focuses on the whole system by measuring effectiveness, efficiency and satisfaction, a much more detailed concept is brought up by Jakob Nielsen [34]. According to Nielsen usability is necessary to gain system acceptability. From this social and practical acceptability can be derived. Usefulness, which is part of the practical side, again includes utility and usability. Figure 1 visualizes this hierarchy and reveals furthermore that the properties Easy to learn, Efficient to use, Easy to remember, Few errors and Subjectively pleasing are requirements for usability. Learnability, which emphasizes that the system should be learned quickly to facilitate users to work with it very fast and quickly, is as well one dimension of usability as efficiency. In contrast to the learning factor efficiency declares how productively users can use a learned system. A further requirement to the system is that it should be easy to remember. Even after a period of time not using it the user should be able to remember how to work with the system. Beyond the explained memorability, a low rate of errors should emerge during use of the system and the user should be able to recover them easily. According to Nielsen satisfaction also is a decisive attribute for usability. In this context it expresses that the user is subjectively satisfied and therefore likes the system [34].

Compared to the ISO definition this view of usability points out, that usability is a quality feature of a product and is independent of context and the person who is using it [7].

Contrary to Nielsens approach, Burmester et al. [7] and Bevan [6] emphasize the importance of context of use. To gain a good quality of a product or system the particular users (for example age, education),

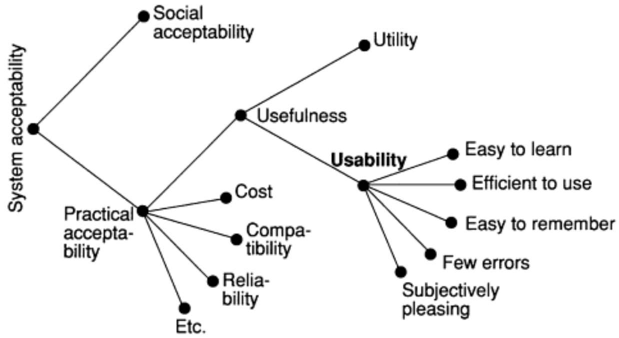


Fig. 1. Nielsen's hierarchy of system acceptability. [34]

tasks (for instance duration or frequency of a task) as well as environmental issues (technical, physical,social,organizational)have to be kept in mind [5, 7]. Therefore these two aspects of usability can be distinguished by their specific focus [6]. Due to the fact of considering context, this paper is based on the definition of the ISO software ergonomics committee.

2.1 Evaluation of Usability

Measuring usability is important to determine whether a product or system is usable or not. Thus, all interactive technologies are affected by usability, which makes it necessary to be considered during the design process as well as to improve systems. For this purpose professionals developed evaluation methods and reliable metrics [12]. The goals of this process can either be to improve the usability of a product as part of design, or to assess the extent to which usability objectives have been achieved [1]. Defining the context is therefore necessary to measure usability.

2.1.1 Context of use

The context of use includes user, tasks, working appliance and physical and social environment in which a product or system is embedded. [19]. In case of users it is important to know their state of knowledge, skills and properties like gender, age and physical abilities. Concerning the tasks the frequency and duration should be known. The functional scopes of hard- and software-components used as working appliance have to be pointed out clearly [19]. Organizational environment, which is one of the three areas derived from the aspect environment, covers the structure of work (e.g. duration of breaks, single work or group work), companys attitude and culture (i.e. business connections) as well as work design (e.g. users freedom of choice, performance measuring or work pace). Configuration of hard- and software is the main goal of the technical aspect of environment. The physical part describes workspace conditions, for example heat or light, workplace arrangements (space, posture) and protective facilities labeled as safety at the workspace [19].

2.1.2 Usability metrics

As mentioned in the definition part, effectiveness, efficiency and satisfaction are attributes of usability. Besides the context of use this aspects have to be specified to make the measuring of usability possible [19]. In this case effectiveness is related to the accuracy and completeness of achieving objectives, whereas efficiency enunciates the ratio of the degree of goal attainment and the effort going with it. Additionally the third aspect satisfaction, which is measured by attitude rating, reveals how comfortable the user is with the use of the system [7, 10].

An example for these aspects would be the scenario of new customers withdrawing money from a cash machine. All these users are novice users (specific users), trying to withdraw a certain amount of money (user goal). 90 percent reaches the goal (effectiveness) in less than one minute (efficiency) with an average satisfaction rating '6' (satisfaction). They can do this operation with any bank machine (context) [22]. In this particular case effectiveness is the amount of users which fulfill their tasks successfully. Another opportunity to quantify this attribute could be to determine fulfillment of an objective percentual. Efficiency, for instance can focus on time or the task itself. In the example shown the focus is set on the task and with it the time the users need to complete this task. By centering time, the accomplished tasks of users during a defined time period are measured. Metrics for satisfaction could be generated by taking complains of the users into account, counting the optional usages or letting users assess the satisfaction by an evaluation scale (shown in the example above) [19]. The necessity of measuring all three usability aspects can be answered by working out the correlation between those aspects.

2.1.3 Correlation between usabiliy aspects

To point out the correlation between the three usability aspects effectiveness, efficiency and satisfaction two studies are presented. Hornbaek and Law [20] conducted a meta-analysis in which they evaluated raw data from usability measures of 73 studies.

The objective was to figure out information about how the measures are related, to understand what the concept of usability is and how to develop models for it. In detail they investigated the relation between effectiveness and satisfaction, efficiency and satisfaction and the relation between specific measures (e.g. task completion time). Further variables that may moderate these relationships (i.e. task complexity) were searched for. The studies used were from eight Human-Computer-Interaction (HCI) journals and conferences. Therefore a broad spectrum of HCI work was considered. They focused on original research papers from the years 2003 to 2005 reporting usability measures concerning human interaction with user Interfaces. Out of these 73 studies 36 measured all three usability aspects and 30 one of the combination of effectiveness-efficiency, effectiveness-satisfaction or efficiency-satisfaction. With only .247 the average correlation between effectiveness and efficiency was weak. In this case 87 percent of the studies had a positive correlation. The highest correlation of a study was .79. Only one study showed a negative correlation. In general there is no significant correlation between those two aspects. Effectiveness and satisfaction correlated up to a value of .164. This average was derived from 39 studies of which 86 percent show a positive correlation. Valued by .196 the correlation of efficiency and satisfaction was measured in 45 studies of which 82 percent showed a positive correlation.

Overall only a medium correlation between those usability aspects exists. The coefficient value ranged between .164 and .247. The complexity of usability measures, prototypical or standard measures as well as measures based on participants perceptions are factors, which were shaping this correlation. Not as expected task complexity seems not to influence the relation.

Similar to the results of Hornbaek and Law , Frokjaer et al. [10] analyzed data from an experiment where 87 subjects solved 20 information retrieval tasks concerning programming problems. They also only found a weak correlation between the aspects effectiveness, efficiency and satisfaction. Therefore they suggest seeing those aspects in general as independent and further submitting that usability testing of computer systems for complex tasks should include efficiency, effectiveness and user satisfaction.

2.1.4 MUSiC

In the following chapter the evaluation-method MUSiC, standing for "Metrics for Usability Standard in Computing", is introduced to give an idea of how evaluation of usability can be carried out. MUSiC gives valid and reliable means of how to measure user performance and satisfaction. Within the frame of this method the term user performance can be subdivided into effectiveness and efficiency [4].

To assess effectiveness two aspects are considered. The quantity of the task the user completes and the quality of the goals the user achieves. Both identifiers are measured as percentage. Whereas quantity is the proportion of the task goals represented in the output of the task, quality shows the degree to which the output achieves the task goals. It can be necessary to measure the effectiveness of sub-tasks, when for example in a drawing task the user is not able to save the final drawing. In this case overall effectiveness would be zero, while some particular sub-task may exhibit a much higher effectiveness. The percentage value of effectiveness can be calculated the following way:

$$Task\ Effectiveness = \frac{1}{100\ x\ (Quantity\ x\ Quality)}\ \% \qquad (1)$$

Efficiency is related to the level of effectiveness achieved proportional to the resources needed. Resources in this context can be mental or physical effort, which makes a statement about human efficiency or time.

$$Temporal\ Efficiency = \frac{Effectiveness}{Task\ Time} \qquad (2)$$

This indicator can be used to compare two or more similar products, types of users or tasks. Conditions for such a comparison are the equality of the environment. Additionally by comparing products the same version of the product has to be used by the same user group for the same tasks. To gain knowledge about the types of users, also the same products and tasks have to be used. Requirements for comparing tasks are the same users and the same product. Comfort and acceptability are the crucial aspects by assessing satisfaction. While comfort is focusing on the emotional part, whether the user feels good or bad by using the system, acceptability emphasizes the users attitude towards the system. This could for instance be how supported the user feels by the system during a particular task [4].

Kirakowski, Porteous and Corbett [26] developed the Software Usability Measurement Inventory (SUMI), which facilitates the measuring of satisfaction. This questionnaire is composed of 50 internationally standardised items, available in several languages. This approximately 10 minute lasting survey requires at least 10 users to get valid results. The results can be provided as and overall assessment and a usability profile which is subdivided in Affect, Efficiency, Helpfulness, Control and Learnability. Another measuring technique suggested in context of MUSiC is to determine cognitive work. During this diagnostic users have to expend excessive mental effort to achieve acceptable performance. For MUSiC two questionnaires are reliable and valid: Subjective Mental Effort Questionnaire (SMEQ) [45] and Task Load Index (TLX) [21]. SMEQ contains only one scale and has the objective to support users to rate the amount of effort invested during task performance [45]. TLX, which is develop by NASA [33], is a multi-dimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales [4].

2.2 Beyond usability

As described in the previews paragraphs, usability is asking about special aspects to be fulfilled. Consequently this model tries to avoid factors like stress and pressure to the user. Beyond this hypothesis factors like Joy of use have to be considered by designing interactive products. Therefore users want to have fun and Joy by using a specific product [8]. Millard et al. [31] showed in an example of the user interface of a call-center software, that especially in working environments an attractive interface, which conveys fun, increases the quality of work [7]. Nevertheless usability is still an important part during the process of designing products or systems. Scott [39] explains with the help of a location-based service, that even if there is more complexity in nowadays products, like the mentioned ”joy of use”, Usability aspects like efficiency and effectiveness are still building an prerequisite for good product design [13]. In the eyes of Dillon [9] usability is important but also insufficient. To him more innovative software is not task-based in the classical performance sense. An example for this could be the processes underlying creative thinking or stimulation, where the cause-effect relationship between input and output are not so clearly delineated. According to a research of Scheldon et al. [42] autonomy, competence, relatedness and self-esteem are the experiential contents and characteristics that make people happiest. To satisfy those psychological needs in the context of product design, the view of satisfaction has to be broadened.

3 Experience Design

Whereas the 20th century is called the ”product age”, nowadays products have the purpose to create an experience [37]. This statement is proven by the reactions happening in leading industries. Phillips for example changed his slogan ”Lets make things better.” by ”Sense and Simplicity”. One year later in 2005, ”Enjoy uncertainty” was the advertising slogan of the iPod shuffle. 2006 Nike launched a champagne named ”Joint Product eXperience” [36]. This theory in change of economy is also confirmed with the theory of Pine and Gilmore [37].

To them the economy changed from being a service-oriented economy towards an economy, which more and more tries to enroll their products with experience. If we take a look at the evolution of a birthday cake this will become clear [37]. Whereas during the agrarian economy mothers made birthday cakes from scratch, in industrial age they draw back on premixed ingredients. Outsourcing the whole process and ordering the cake from a bakery was the next step involved in service economy. Today parents not buy a birthday cake or throw a party, what they do is to outsource the whole event to a company that stages a memorable event for the kids. With this four-stage example Pine and Gilmore emphasize the importance of experience in nowadays economy. Another advantage that lies in designing for an experience is the interdisciplinarity. Ideas from disciplines like psychology, communications, and computer science can be combined [28].

In the area of Human-Computer-Interaction the term user experience is of interest. Although it does not much differ from the normal experience the focus is laid on interactive products [16]. Figure 2 shows the scope of user experience. In detail it focuses on products, systems, services and objects that a person interacts with through a user interface [27].

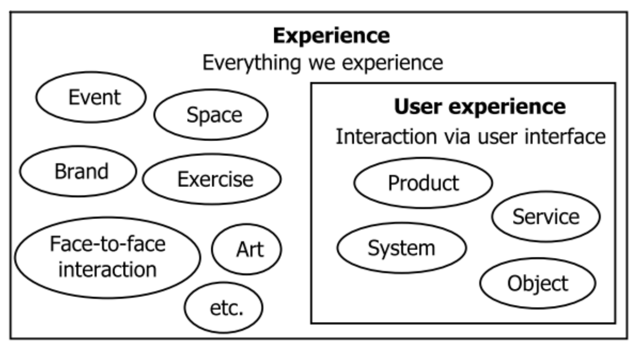


Fig. 2. The scope of experience and user experience [27].

3.1 Defining User experience

According to The User Experience Professionals’ Association user experience is defined as

> ”Every aspect of the user’s interaction with a product, service, or company that make up the user’s perceptions of the whole. User experience design as a discipline is concerned with all the elements that together make up the interface including layout, visual design, text, brand, sound, and interaction. User Experience works to coordinate these elements to allow for the best possible interaction by users.” [1]

The ISO software ergonomics committee expects user experience to be

> ”aspects of the users experience when interacting with the product, service, environment or facility [...] It includes all aspects of usability and desirability of a product, system or service from the users perspective.” [14]

This definition seems to have two inconsistency. At first the term all aspects is a very broad and open variable, which can be compared to an empty shell. It can be filled with everything. Due to this vague formulation debates are avoided but are not helpful. Desirability is the second term, which is also not well defined and leads to more questions than answers. A definition by Hassenzahl [14] subdivides user experience into two parts. The first part defines user experience itself. It is

> "[...] a momentary, primarily evaluative feeling (good-bad) while interacting with a product or service." [14]

Further

> "[...] good user experience is the consequence of fulfilling the human needs for autonomy, competency, stimulation(self-oriented), relatedness and popularity (others-oriented) through interacting with the product or service. Pragmatic quality facilitates the potential fulfillment of be-goals." [14]

To understand this definition, it should be clear how an experience arises. This is explained in the following chapters.

3.2 Users needs

In 2001 Sheldon et al. [42] compared ten candidate psychological needs with the objective of determining which of this needs are most fundamental for humans. This was done with the aid of three studies. Therefore the participants described most satisfying events within their lives and then rated the salience of each of the ten candidate needs within these events. Autonomy, competence, relatedness, physical, security, self-esteem, pleasurable- stimulation, money-luxury, popularity-influence and self-actualization are the examined needs proposed by prominent psychological theories. Their results showed that autonomy, competence, relatedness and self-esteem are experiential contents and characteristics, which make people happiest [42].

Another more technologically oriented study made by Hassenzahl and Diefenbach [14] determined that experienced autonomy and competence in context of using technology is a source of positive experience. To figure this out they asked 52 persons to think of a recent, very positive satisfactory experience with a technology. The participants had to specify their answers, by questionnaire. It included nine items to measure the chosen needs autonomy, competence and relatedness. Whereas autonomy and competence are linked to positive experience there is no significant correlation of relatedness to positive affect.

These fundamental types of user experience (relatedness, competence and autonomy) have to be considered by creating a product, which has the claim to create positive user experiences. To ease the process of creating such products, an approach developed by Mark Hassenzahl subdivided the quality of a product into two parts.

3.3 Hedonic and pragmatic quality

According to the two-component-model [18], two types of quality can be distinguished by people perceiving a product.

The pragmatic quality implicates the utility and usability of a product. Thereby the focus lays on the product self and implicates human needs like safety, control and trust. Pragmatic quality is needed to achieve do-goals. Do-goals can for example be making a telephone call or finding a book in an online store.

In contrast be-goals can be fulfilled by hedonic quality. Perception of non goal-oriented characteristics of quality is the focus of hedonic quality. Such goals can be being competent, being related to others. It answers the question why a user owns a product. Human needs like self-expression, novelty and change are related to hedonic quality [7, 14]. In the year 2000 Sandweg, Hassenzahl und Kuhn made an empirical evaluation about a telephone based user interface for controlling a house-automation-system. During this evaluation a high pragmatic but low hedonic and aesthetic value of the user interface was measured. The low value was due to the fact that a voice input was just to interact with the user interface and all other non-voice sounds were missing. As know in human-computer-interaction this kind of voice can have a positive effect towards interaction. This can be compared to two documents which one of them includes illustrations and the other one only consist of blank text. Thanks to the measurement of hedonic quality it was possible to remedy this deficiency [7].

An example for a hedonic focused design is the bag of knowledge developed by Mark Hassenzahl [15]. In this bag a mp3-player is embedded, which plays pre-recorded snippets of psychological knowledge. By designing the bag he focuses on particular experiences around the themes surprise, curiosity and competence. Therefore this bag should be placed in a surprising location, such as supermarket shelves, next to regular products. Further the user has no control of the sequence played by the mp3-player. It is playing the tracks continuously and it is not possible to manipulate this procedure. Another requirement is that user can only consume the audio tracks individually. Because of this the volume of the tracks is so low that the user has to bring the ear close to the neck of the bag. With the help of the low voice and the fact that the user has to listen in the bag Hassenzahl create the experience of surprise and curiosity. To gain experience of competence he created the impression of being told secrets, to be shared immediately or stored away for later use. With this example Hassenzahl showed a way to create a product by focusing on be-goals such as surprise and curiosity for the design process.

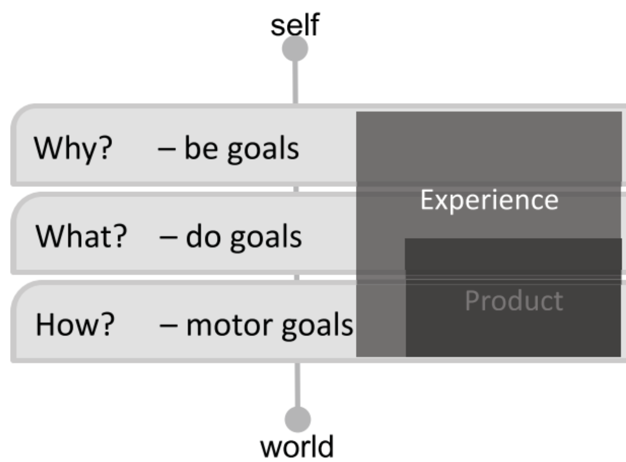


Fig. 3. The three types of goals an experience consists of (be-, do- and motor- goals). [15]

3.4 Visceral, Behavioral and Reflective Level

In contrast to animals, humans are able to know their roles in the world and can reflect upon past experience [35]. Figure 4 shows the three levels from which human attributes result. Imaging putting a wire fence mesh and some desirable food. A chicken would likely be stuck in the fence (visceral). Whereas the dog runs around the fence each time, human beings have not only the ability to realize that they have to run around the fence (behavioral), they also have the cognitive skills to reflect the situation and move the fence or the food so that they do not have to run around the fence. Further humans can communicate to others and reflect the situation (reflective) [35].

3.4.1 Visceral

This level is called the automatic or prewired layer. It is pre-consciousness or pre-thought. The brain analyzes the world and responds. An example for visceral level is riding a rollercoaster [35]. During this process people are influenced by falling, excessive speed and heights. To design for the visceral level means to focus on the initial impact of a product. The appearance of the product matters and with it how it touches and feels. Contrary to the reflective level, where consciousness and interpretation is possible, in visceral level there is only affect [35].

3.4.2 Behavioral level

This level allows analyzing a situation and altering behavior accordingly. Precondition is a complex and powerful brain work. One property of behavioral level is, that it is not conscious. This is why for example piano player are able to let their fingers play automatically while they were having a conversation [35]. Behavioral level can also be the pleasure any expert feels when doing something well. This can be driving a difficult course or playing a complex piece of music. Within the framework of the design process, the behavioral level is focused on the use of a product. Performance, function and usability are the main aspects have to be considered. By fulfilling this attributes of the behavioral level the user gains warm and positive affects when using the product. The visceral as well as the behavioral level are about the feelings and experiences while actually seeing or using a product [35].

3.4.3 Reflective level

Named at the highest evolutionary level, the reflective level gives Human Beings the opportunity to think about their operations [35]. Due to this conscious thinking, learning of new concepts and generalization about the world is possible. Reflective brainwork for instance means contemplating a serious literature or art. To gain enjoyment this requires study and interpretation. On this level therefore consciousness and the highest levels of feelings, emotions and cognition reside. It further facilitates understanding and interpretation. It can easily be influences by culture, experience, education and individual differences. Unlike behavioral and visceral the reflective level extends much longer and is not limited to the experience while using the product. Because of reflection it is possible to remember the past and contemplate the future. By designing for the reflective level aspects like self-image of the user, personal satisfaction and memories have to be considered. Long-term relations as well the feelings produced by owning the product have to be kept in mind. Also identity plays a major role. It can be expressed in pride of ownership or use [35].

Even though the individual levels are characterized clearly the way they interact to each other a complex. The highest level (reflective), can for example override the other ones. Further any real experience involves all three levels [35].

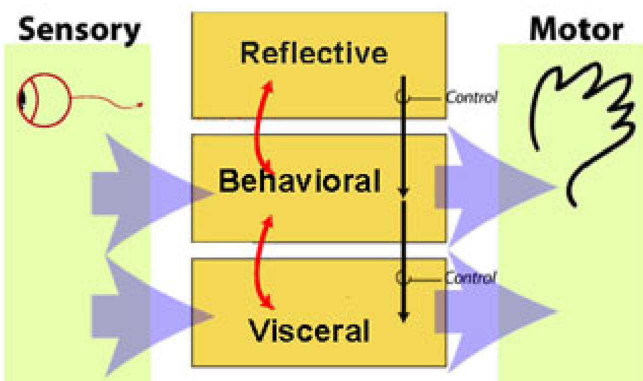


Fig. 4. The three levels from which human attributes result (Visceral, Behavioral and Reflective). [35]

3.5 Duration of an experience

Although each experience is a unique artifact, the term an experience is not restricted to the moment of actual usage. People can also have indirect experience before and after usage [38].

As shown in figure 5 the so called anticipated experience reflects the expectations a user has in mind before using the product or system. This can for example be an experience from related technology, the image of a brand or a particular advertisement [38]. While in the episodic phase the user reflected and evaluates the currently experienced situation, momentary user experience is the current state of having an experience. Focusing on this type of experience can for instance be helpful if you want to gather information about a persons emotional response to details of a user interface. Cumulative user experience gives a whole view of the system. Such a type of experience is characterized by having periods of non-use and episodes of usage. Within the scope of cumulative, which can be structured to a lifecycle, user experience it is possible that pervious experience may influence a future one, by for example an episodic becoming an anticipated user experience [38].

3.6 Properties of experience

Subjective, holistic, situated and dynamic are the properties representing experience [15].

3.6.1 Subjective

Due to the fact that an experience is created in the head of the experiencer, experience is subjective. Therefore not only the objectivity of a product matters, it also must be able to be experienced [15]. To point this out Jones and Peppiat [23] showed in there study about waiting time in queues that occupied time is perceived as shorter than unoccupied time. The wait fell shorter and closer to reality, given a running television set was present as a distractor. [15] Another example is the difference between winning metals during the Olympic games. At the Summer Olympics in 1992 Medvec et al. [29] examined the emotions and reactions of athletes winning bronze and silver metals. What they figured out was that even though both did not win a gold metal the perceived experienced differs. It turned out that the athlete winning bronze is happier than the bronze medalist. This can be explained by the downward and upward counterfactual. Whereas for the bonze-winning athlete the most compelling alternative is to finish without a metal (downward), the most compelling alternative is to the silver-winning athlete is to win gold. This fact of downward and upward thinking is explained by Norm Theory [25]. To compare reality people always look for alternatives (fantasy). During this comparison process people start to think about hypothetical cases and therefore take the nearest miss. For are bronze-wining athlete this is not wining a metal, for a silver-wining athlete this is not winning the gold metal [15].

3.6.2 Holistic

As mentioned before, there are different goals, which have to be distinguished by designing for an experience [15]. Holistic in this case means, to consider all of this goals. As shown in Figure 3, do-goals are on the middle level. They focus on question what to do. It is a concrete outcome and can for example be making a telephone call or watching a movie [15]. Although such kind of goal is not completely inside the technological context, it is at least born out of a general technology. It is possible to watch a movie with the help of a variety range of devices (Television, Tablet, PlayStation and so forth), but technology is still necessary to watch the movie.

The lowest level of goals is called the motor-goals [15]. This low goals concentrate on the how. They are derived from the higher do-level. For instance if you make an telephone call on your IPhone you first have choose a green icon with the label phone. Next a screen pops up and you have to select the menu item keypad in a bottom-oriented menu, which brings up a user interface with numbers to dial on it. After you dialed your desired number you have to push a digital green button with a telephone receiver on it. As shown in the example the do-goal of making a telephone call is subdivided into several motor-goals.

Be-goals are the third type of goals. They motivate action and provide them with meaning. By nature they are self-referential and therefore close to peoples selves [15]. A be-goal can for instance be being competent, being close to others, being autonomous or being stimulated. When making a telephone call, the telephone call itself is not meaningful. But in a particular context this call can get of high meaning of people. Imaging the situation when you missing a flight and have to spend a night in a city you do not know and further do not have any friends in this town. In this case experiencing loneliness is

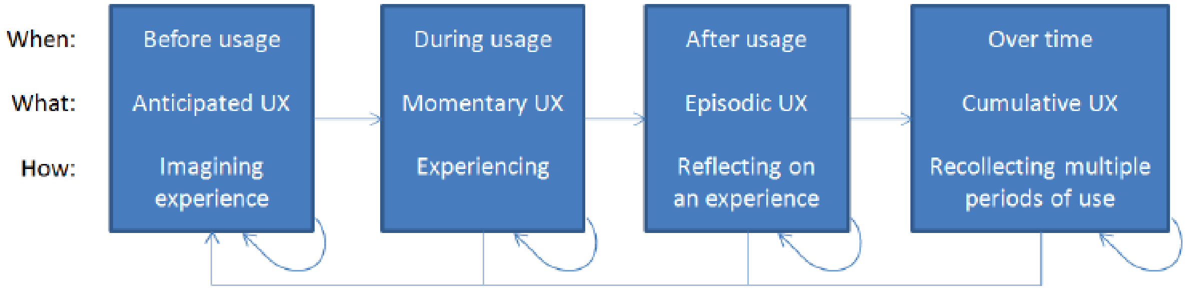


Fig. 5. The different time spans a user experience can be subdivided into. [38]

very likely. To fulfill the need of being close to others in this situation you may make a telephone call. The call becomes meaningful.

Hence, to design and evaluate experience implies take you have to take in mind all three levels motor- and do- as well as be-goals [15].

3.6.3 Situated

Looking at an experience shows that it is a highly situated, idiosyncratic, emerging entity. One experience is never alike another neither between nor within experientors [15]. It is a unique combination of perception, motivation and emotion during a particular time at a specific place. Think of the situation a man is writing a lovely SMS to his wife. The man of course writes a lot of SMS in his life. But this particular SMS will compose a specific experience that cannot be recovered in the same way. Hence, to design an experience means to take in mind that each experience is highly situated in itself. [15].

3.6.4 Dynamic

Experience can change. Not only depending on the situation or the context, it can also change in itself. Experience is a continuous stream emerging from perceiving, acting, thinking and feeling. In other words it can be a chunk of time, packaged, interpreted and labeled. Hassenzahl suggested to distinguish in the field of Human-Computer-Interaction between micro, meso and marco perspective in case of the dynamic of experience. Whereas micro only covers a short time span of hours, meso experience measurement is based on duration of several weeks. However, marco with a scope on years of use. [15] To make the dynamic of experience clear, several studies about different periods of times will be introduced.

During a period of eight weeks Mendoza and Novick[30] were observing participants about their level of frustration using software for creating websites. They found out that during the eight weeks the main effect was that the frustration for using the software was dropping. Further they determined that users during the episode of experience get more autonomic by solving problems on their own instead of asking other people. Due to this fact the way they experienced the same tasked changed. This kind of study explaining the dynamic of experience can be categorized as meso-oriented.

In a macro-oriented study, Wilamowitz-Moellendorff et al. [44] wanted to point out, if there is a change in the perception of users towards a particular product during a specific time episode. The average time span was 22 month. 57 people were interviewed, 20 about their mobile phone, 18 radiographers about their used Computer Tomography scanner and 19 people about productivity software like Mircosofts excel or word. Participants had to give feedback about the usage of now until the beginning of the relationship with the product. Utility (provides the product all necessary functions), Usability (are the functions provided in an easy and efficient way), stimulation (can the product surprise or foster the curiosity and provide the opportunities for the perfection of knowledge and skills), Beauty (is the product able to evoke beauty) and Communicate identity (ability a product communicates self-serving symbols to relevant others) are the several dimensions on which the participants had to describe their experiences.

Whereas utility (UT) and usability (US) can be categorized as pragmatic quality, stimulation (S), Communicate identity (I) and beauty (B) represent hedonic quality [44]. Figure 6 provides an overview of the results. As can be seen the dynamic between the three product types is highly differing. Within the specific type of product participants experienced more or less the same. Therefore designing of the features specific types of products have, can probably be a good indicator designing for. Further the study determined that the experience products provide change during time of use. In the beginning a mobile phone is a kind of device, which is stimulating and demonstrates beauty. After a while the proud of the owner towards the mobile phone decreases and the phone turns into something utterly pragmatic (Figure 6 that all hedonic features were deteriorating during time). In contrast to this due to the complexity of the CT scanner, it needs more time to be mastered. Usability therefore accelerates improvingly. Further over time the complexity participants got used to stimulate them to think about new opportunities in context of using the product [15].

These studies point out the dynamic nature of experience.

	Dimension				
Product	UT	US	S	B	I
Mobile	Stable	Accelerated improvement	Accelerated deterioration	Jumping deterioration	Steady deterioration
CT	Stable	Accelerated improvement	Accelerated improvement	Stable	Stable
Productivity	Steady improvement	Steady improvement	Steady improvement	Stable	Steady improvement

Fig. 6. Results of how the perceptions of users change towards a specific product during a specific time episode. [44]

4 Relation between Usability and User Experience

There is no common definition of the relation between usability and user experience. One approach is to see usability as precondition for User Experience. As a consequence usability is part of user experience [24, 32, 43].

This theory is challenged by Hassenzahl [15]. He explains with an example of a souvenir, that these two qualities of a product are independent. A souvenir such as the gondola from Venice provides a specific experience by reflecting the happenings of the visit. Notwithstanding this figure has a high hedonic value, whereas there is no pragmatic quality. He also makes clear that in the area of Human-Computer-Interaction the focus mainly lays on fulfilling do-goals, such as making a telephone call and the be-goals, which create positive experiences during such a process, are excluded. To him neither products which provide function without meaning, nor those which provide meaning without function are likely to be successfully.

In contrast to the first theory there is another argumentation saying that user experience is a broad and rich term for satisfaction. According to this, user experience is part of usability due to the fact that the users' satisfaction is an objective to be fulfilled in context of usability [2, 32].

As shown, there is no common theory, which structures or categorizes the relation between user experience and usability in a satisfying manner. Fact is that user experience and usability are indicators of the quality of a product and should be considered by designing products.

Further it can be useful neglecting usability requirements when designing for a particular experience. One example is a mp3-player concept developed by students of the Follkwang Hochschule [17]. The main goal of this concept is to experience music together. Therefore an mp3-player is loaded with favorite music before meeting friends. This player only plays the music when surrounded by other players. When players are next to each other a mixed playlist is played. With the conscious omission of functionality like letting the mp3-player play on his own or having a display to see which song is played (usability), the concept strengthens the experience gained by using the product. A second example for neglecting usability requirements for experience centered design is the DriftTable designed by Gaver et al. [11]. It is a coffee table with a peephole in its middle. When looking through the peephole an aerial image of Great Britain is shown. The journey starts above your own residence and by giving pressure to the table in cardinal direction you want to travel, the journey can be controlled at a maximum speed of 50 km/h. By using this concept functionality is limited too. The designers force the user to gain the experience, by giving a restricted variety of functional opportunities [17].

5 Conclusion

The first part of this paper gives a broad overview of the term usability. Three definitions are introduced and compared. These definitions have a different focus. Whereas the ISO software ergonomics committee [7] points out that the objective of usability is to create effective, efficient and satisfying products with the consideration of the user and the context of use, the approach of Nielsen [34] explains usability as a quality feature which is independent from the context of use and the user.

I recommend regarding usability as a quality aspect, which depends on the user and the context of use. This is due to the fact that nowadays a broad range of devices (i.e. tablets, laptops and smartphones) can be used to reach a specific goal (e.g. open a website), but the context and the user will always be decisive factors for creating usability. Therefore the definition of the ISO software ergonomics committee should be taken in mind when considering the term usability. Hornbaek and Law [20] and Frokjaer et al. [10] demonstrated the weak relation between the aspects effectiveness, efficiency and satisfaction. This finding additionally emphasizes the strength of the ISO definition.

A further question, which the paper gives an answer to, is the approach of experience design. This developing process focuses on the user's needs and simultaneously points out aspects which have to be considered beyond usability. It is also important to distinguish the terms experience and user experience. Whereas experience covers everything we experience, user experience focuses on the way we interact with products, systems, objects or services via user interfaces. To ease the process of designing for a specific user experience I recommend to subdivide product quality into hedonic and pragmatic quality. This two-component-model created by Mark Hassenzahl [7] emphasizes that besides usability (pragmatic quality) there is a further aspect which should be considered. Aspects like autonomy, relatedness or competence are considered.

In the third part of this paper the relation between the concepts usability and user experience is elaborated. There is no common point of view concerning this issue. One opinion is to regard experience as a broadened term of satisfaction and therefore as a part of usability. Another point of view claims that usability is a precondition for user experience. In this paper the latter approach is supported. This is due to the fact that when designing a product for a specific experience, usability always has to be kept in mind. Even though usability aspects are neglected (as shown by two examples in the paper) it is necessary to be conscious about them.

References

[1] U. P. Association. Glossary - usabilitybok.org. `http://usabilitybok.org/glossary/19\#term432`. Last accessed: 09-03-14.

[2] S. Benford, G. Giannachi, B. Koleva, and T. Rodden. From interaction to trajectories: Designing coherent journeys through user experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 709–718, New York, NY, USA, 2009. ACM.

[3] J. Bennett. Managing to meet usability requirements: Establishing and meeting software development goals. *Bennett, J.L., Case, D., Sandelin, J., Smith, M. (Eds.), Visual Display Terminals: Usability Issues and Health Concerns*, pages 161–184, 1984.

[4] N. Bevan. Measuring usability as quality of use. *Software Quality Journal*, 4(2):115–130, 1995.

[5] N. Bevan. Usability is quality of use. *Advances in Human Factors Ergonomics*, 20(B):349–354, 1995.

[6] N. Bevan. Quality and usability: a new framework. *Achieving software product quality, van Veenendaal, E, and McMullan, J (eds)*, pages 25–34, 1997.

[7] M. Burmester, M. Hassenzahl, and F. Koller. Usability ist nicht alles wege zu attraktiven produkten (beyond usability appeal of interactive products). *I-Com Zeitschrift für interaktive und kooperative Medien*, 1(1/2002):32–40, 2002.

[8] J. M. Carroll and J. C. Thomas. Fun. *ACM SIGCHI Bulletin*, 19(3):21–24, Jan. 1988.

[9] A. Dillon. Beyond usability: process, outcome and affect in human-computer interactions. *Canadian Journal of Library and Information Science*, (March):57–69, 2002.

[10] E. Frøkjær, M. Hertzum, and K. Hornbæk. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 345–352, New York, NY, USA, 2000. ACM.

[11] W. W. Gaver, J. Bowers, A. Boucher, H. Gellerson, S. Pennington, A. Schmidt, A. Steed, N. Villars, and B. Walker. The drift table: Designing for ludic engagement. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pages 885–900, New York, NY, USA, 2004. ACM.

[12] C. Gilbert. Usability Evaluation. `http://www.interaction-design.org/encyclopedia/usability\_evaluation.html\#heading\_The\_Origins\_of\_HCI\_and\_Usability\_html\_pages\_12447`. Last accessed: 09-03-14.

[13] M. Hassenzahl. The Effect of Perceived Hedonic Quality on Product Appealingness. *International Journal of Human-Computer Interaction*, 13(4):481–499, Dec. 2001.

[14] M. Hassenzahl. User experience (ux): Towards an experiential perspective on product quality. In *Proceedings of the 20th International Conference of the Association Francophone D'Interaction Homme-Machine*, IHM '08, pages 11–15, New York, NY, USA, 2008. ACM.

[15] M. Hassenzahl. *Experience Design - Technology for All the Right Reasons*. Morgan Claypool Publishers, 2010.

[16] M. Hassenzahl, S. Diefenbach, and A. Göritz. Needs, affect, and interactive products Facets of user experience. *Interacting with Computers*, 22(5):353–362, Sept. 2010.

[17] M. Hassenzahl, K. Eckoldt, and M. T. Thielsch. User Experience und Experience DesignKonzepte und Herausforderungen. *Usability Professionals*, pages 233–237, 2009.

[18] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner. Hedonic and ergonomic quality aspects determine a software's appeal. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00*, 2(1):201–208, 2000.

[19] A. M. Heinecke. *Mensch-Computer-Interaktion - Basiswissen für Entwickler und Gestalter*. Springer-Verlag, Berlin, Heidelberg, New York, 2. aufl. edition, 2012.

[20] K. Hornband E. L.-C. Law. Meta-analysis of correlations among usability measures. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, page 617, 2007.

[21] E. Houwing, M. Wiethoff, A. Arnold, and E. Vos. *Introduction to Cognitive Workload Measurement*. Laboratory for Work & Interaction Tech-

nology, Delft University of Technology, 1993.

[22] T. Jokela, N. Iivari, J. Matero, and M. Karukka. The standard of user-centered design and the standard definition of usability: Analyzing iso 13407 against iso 9241-11. In *Proceedings of the Latin American Conference on Human-computer Interaction*, CLIHC '03, pages 53–60, New York, NY, USA, 2003. ACM.

[23] P. Jones and E. Peppiatt. Managing perceptions of waiting times in service queues. *International Journal of Service Industry Management*, 7(5):47–61, 1996.

[24] P. W. Jordan. *Designing Pleasurable Products - An Introduction to the New Human Factors*. CRC Press, Boca Raton, Fla, 2003.

[25] D. Kahneman and D. t. Miller. Norm theory: Comparing reality to its alternatives. *Psychological Review*, pages 136–153, 1986.

[26] J. Kirakowski and M. Corbett. Sumi: the software usability measurement inventory. *British Journal of Educational Technology*, 24(3):210–212, 1993.

[27] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort. Understanding, scoping and defining user experience: A survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 719–728, New York, NY, USA, 2009. ACM.

[28] H. McLellan. Experience design. *Cyberpsy., Behavior, and Soc. Networking*, 3(1):59–69, 2000.

[29] V. H. Medvec, S. F. Madey, and T. Gilovich. When less is more: counterfactual thinking and satisfaction among Olympic medalists. *Journal of personality and social psychology*, 69(4):603–610, Oct. 1995.

[30] V. Mendoza and D. G. Novick. Usability over time. In *Proceedings of the 23rd Annual International Conference on Design of Communication: Documenting and Designing for Pervasive Information*, SIGDOC '05, pages 151–158, New York, NY, USA, 2005. ACM.

[31] N. Millard, L. Hole, and S. Crowle. Smiling through: Motivation at the user interface. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I - Volume I*, pages 824–828, Hillsdale, NJ, USA, 1999. L. Erlbaum Associates Inc.

[32] I. M. Moczarny, M. R. R. de Villiers, and J. A. J. van Biljon. How can usability contribute to user experience?: A study in the domain of e-commerce. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, SAICSIT '12, pages 216–225, New York, NY, USA, 2012. ACM.

[33] H. P. G. NASA-Ames Research Center. Collecting NASA Workload Ratings: A Paper-and-Pencil Package. Moffet Field, CA: NASA-Ames Research Center. 1986.

[34] J. Nielsen. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[35] D. A. Norman. *Emotional Design - Why We Love (or Hate) Everyday Things*. Basic Books, United States, 2007.

[36] E.-J. R. G. Oppelaar, E.-J. Hennipman, and G. C. van der Veer. Experience design for dummies. In *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction*, ECCE '08, pages 27:1–27:8, New York, NY, USA, 2008. ACM.

[37] B. J. Pine and H. J. Gilmore. Welcome to the Experience Economy. *Harvard business review*, (August):97–105, 1998.

[38] V. Roto, E. Law, A. Vermeeren, and J. Hoonhout. User experience white paper. *Bringing clarity to the concept of user experience*, 2011.

[39] K. M. Scott. Feature: Is usability obsolete? *interactions*, 16(3):6–11, May 2009.

[40] B. Shackel. The concept of usability. proceedings of ibm software and information usability symposium. *Bennett, J.L., Case, D., Sandelin, J., Smith, M. (Eds.), Visual Display Terminals: Usability Issues and Health Concerns*, pages 45–88, 1981.

[41] B. Shackel. Usability Context, framework, definition, design and evaluation. *Interacting with Computers*, 21(5-6):339–346, Dec. 2009.

[42] K. Sheldon and A. Elliot. What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of personality and social psychology*, 80(2):325–339, 2001.

[43] H. Väätäjä, T. Koponen, and V. Roto. Developing practical tools for user experience evaluation: A case from mobile news journalism. In *European Conference on Cognitive Ergonomics: Designing Beyond the Product — Understanding Activity and User Experience in Ubiquitous Environments*, ECCE '09, pages 23:1–23:8, VTT, Finland, Finland, 2009. VTT Technical Research Centre of Finland.

[44] M. von Wilamowitz-Moellendorff, M. Hassenzahl, and A. Platz. Veränderung in der wahrnehmung und bewertung interaktiver produkte. In T. Gross, editor, *Mensch und Computer 2007: Interaktion im Plural*, pages 49–58, München, 2007. Oldenbourg Verlag.

[45] F. Zijlstra. *Efficiency in work behaviour: a design approach for modern tools*. Dissertation, Delft University of Technology, November 1993.

Experience Prototyping: From the Ideas to the Product

Luisa Wurm

Abstract— ”Prototypes” are illustrations of a design. They are made before final products are in existence [4]. Prototyping is an important step to develop interactive systems. Thereby the focus of Experience Prototyping is on the users' experiences. The goal of User Experience is to let the user have sensations, emotions and experiences with the product. But how can we design such products? In this paper the terms Usability, User Experience, experience and emotion are defined and their particular context is explained. After that the steps before creating prototypes are listed, for example how to get ideas for new products. Also storytelling is an important technique. Afterwards the different characteristics of prototypes are distinguished and explained. Then User Experience Evaluation Methods are presented and after that a discussion is led about the differences in the process from ideas to prototypes and then to the evaluation between Usability and User Experience. This paper gives answer to five questions. What is the meaning of Usability, User Experience and experiences? What is to do before developing prototypes? How to create prototypes? What are the User Experience Evaluation Methods and the last question is what are the differences and similarities between Usability and User Experience in view of the development process?

Index Terms—Usability, User Experience, Experience, Prototype, evaluation methods

1 Introduction

There is a habit in Italy. If you call a friend, but put down the phone before your friend picks up, it means: ’I just thought of you’. This habit is called squillo [25]. Calling a person but putting down the phone, before he can answer is not really the functionality of a phone. But this gesture associates positive feelings, emotions and the people get new experience with a product. These characteristics are important for User Experience (UX). But what is the meaning of this term?
User Experience is often used in human computer interaction, but is often not clearly defined. Frequently UX is treated as equal with Usability [26]. Looking at one of the popular definitions from ISO 9241-11 [21], Usability is defined as:

> ”The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [21].

With this in view the scope is the result of using a product, measured by [3]:

- effectiveness: this dimension is achieved, when the aspirations of use are reached
- efficiency: describes the resources that have to be expended to reach the aspirations
- satisfaction: this dimension is achieved, when the user thinks that the use of this product is acceptable [3]

Therefore Usability deals primary with the user's cognition and performance in interactions, whereas User Experience handles with non utilitarian aspects of interactions. That means that the focus is on the user affect, the sensation and the meaning just as the value of interaction [26]. The following section deals with different definitions of User Experience. They are compared to each other and concepts are extrapolated to develop specific Experience Prototypes.

2 Defining User Experience

Because of several definitions of User Experience, the term has many variant ways to be defined. Each single approach shows a concept from a different point of view [34]. But why is it so complicated to get a usable definition of UX? One reason is that there are a lot of concepts, including emotional, experiential, aesthetic and many other variables [26]. Another reason is that it is difficult to limit the term. On the one hand there is a single aspect of an end user's interaction with a standalone application and on the other hand there are all aspects of several end users' interactions with the company and its service [26]. This section is concerned with different definitions of UX. The definitions are compared to each other and at last a result is drawn out of that.

The first definition is from J. Nielsen and D. Norman [30] and they define User Experience as follows:

> ”*”User experience”* encompasses all aspects of the end-user's interaction with the company, its services, and its products” [30].

They want to state that experience already starts, when a service of a company is used. Using a service means that the user gains experience when using it, because of the aesthetic and the gained emotions. One example is car sharing where people can rent cars for a certain period. There are already emotions when interacting with the service, for example the user is excited and looks forward to the trip before it starts.
Looking at the definitions from ISO 9241-11 [10], UX is defined as:

> ”A person's perceptions and responses that result from the use and/or anticipated use of a product, system or service” [10].

The focus of this definition lies on the direct result of use, the perceptions and responses. But this definition also declares that UX is related to the product, system or service. The company itself is not mentioned in this definition.
The next definition is from M. Hassenzahl and N. Tractinsky [18] and they define UX as:

> ”A consequence of a user's internal state, the characteristics of the designed system, and the context within which the interaction occurs” [18].

The user's internal state is concretized as assessments, needs, mood, expectations and motivation [31]. The characteristics of the designed system are specified as usability, intention, functionality and complexity. The context can be equaled with the environment. That means that it can be for example a social or organizational environment, the significance of the activity or the gratuitousness of use [31]. E. Law et al. [26] tried to understand, scope and define User Experience in

- *Luisa Wurm is studying Media Informatics at the University of Munich, Germany, E-mail: l.wurm@campus.lmu.de*
- *This research paper was written for the Media Informatics Advanced Seminar ’Media Informatics’, 2014*

their paper. They collected the views of UX from 275 researchers and practitioners from academia and industry. After comparing, evaluating and scoping them, they came to this conclusion:

> ”In summary, we recommend the term user experience to be scoped to products, systems, services, and objects that a person interacts with through a user interface. [...] According to our views, user experience focuses on interaction between a person and something that has a user interface” [26].

They specify the user interface for example as tools, entertainment services or knowledge systems. In their opinion, if there is no interaction with a user interface, there is no User Experience [26].

J. McCarthy and P. Wright [28] provide in their book foundations for a plainer analysis of User Experience in view of technology as experience. Their starting point is the pragmatism of philosophers of User Experience. They especially consider the philosophers John Dewey and Mikhail Bakhtin:

> ”For Dewey, experience is constituted by the relationship between self and object, where the self is always already engaged and comes to every situation with personal interests and ideologies. [...] For Bakhtin, the unity of felt experience and the meaning made of it are never available a priori but must always be accomplished dialogically” [28].

McCarthy et al. focus on the holistic nature of an experience. That means, it is important how meaning is made of it. The approach is more interpretive and qualitative towards user experience [12].

In summary all these definitions have resemblances. All have the focus on the interaction of the user and the resulting factors. Sometimes the definitions differ in the context. I personally prefer the definition of the ISO Standard, because the focus is on the perceptions and the responses of use and the context depends on the product, system or service. I think this definition is easy to understand and refers to the usage. The results of interacting with a system, service or product are the experiences and emotions. In the following section these terms are defined.

3 Defining experiences and emotions

User Experience is not equatable to Usability related to the definitions and in UX the term ”experience” plays an important role. To show the importance of experience, two different things are compared:

- a concert ticket of your favourite star
- a new cool t-shirt

Both are worth 50 euro. Which of this both things makes us happier [17]? It is the first one. Boven and Gilovich [6] show in several studies that the experience makes us happier than products and they look like more worth spending money on it. Hassenzahl et al. [17] describe that experiences are subjective according to their definition. They take place in the ’head’ of the user. That means experience is a psychological phenomenon.

McCarthy and Wright [28] name four threads of experience: the sensual, the emotional, the compositional and the spatio temporal. These threads are ideas to think more obvious about technology as experience. The sensual thread of experience deals with our sensory engagement with a situation. If the functions of the senses are completely implemented to make sense of the situation, the interaction between the user and the environment becomes involvement and communication. The emotional thread relates to the value of judgments. This value is reduced to other people and the importance of things in the view of our needs and desires [28]. Therefore our emotions change if the relation with people or things modify [22]. As a consequence emotions have a motivational function. They prioritize the different actions by giving it a sense of precedence [22]. The compositional thread deals with the relationship between parts and the whole of an experience [28]. For example if you look at a painting, you see the whole painting and the composition of different elements of the painting and their effect. The last thread is the spatio-temporal thread. Interactions build experiences of space and time. Time can be increased or decreased, speed can be slowed down or speeded up and spaces can be opened up or closed down. Space and time can be connected or disconnected [28].

The sensual and the emotional thread appreciate centrality of the senses and feelings, the compositional and spatio temporal thread appreciate indivisibility of an experience and especially its dynamics [17]. Experiences attribute a new meaning to our acting. They remember, communicate and influence the motivation (positive or negative). Interactive products can play a role in this experiences: as a trigger or multiplier of experience [17].

A smart phone for example makes us independent, connects us with our friends or represents style [17]. The sensual thread connects the user’s sensory with the experience. The user can for example feel comfortable when using the smart phone or he can feel uncomfortable if the smart phone’s handling is complicated. The emotional thread changes if the user does an action. If the user for example reads a nice text message from a friend he feels happy and he is motivated to write something back. The compositional thread deals for example with the relationship between the experience of the different parts of the smart phone and the experience of the smart phone itself. There are applications that cause positive and there are applications that cause negative experiences. These influence the overall experience with the smart phone. The spatio-temporal thread declares that interactions build experience of space and time [28]. If the user is in a hurry for example the space of the smart phone may appear to small, because the small screen can not show all the information the user wants to see at the moment. If the user is not in a hurry the space does not disturb because he has the time to perform more than one interaction. These experiences make a product significant and generate a bonding between the user and the smart phone [17].

4 How to generate ideas

It is important to understand the single meanings of User Experience, experience and emotion and their correlation before thinking about developing new products, because this comprehension influences the idea generating process. Before creating prototypes you have to get an idea what product you want to develop. Anu Kankainen [22] describes three approaches for product development: conventional continuous product development, explorative product development and User Centered Product Design. The conventional continuous product development is analysis goaded and it appeals at progressive changes in product families [22]. The techniques in the early phases of the product development process are for example concept tests, focus groups or conjoint analysis. With the help of these techniques, the development team tries to get answer to questions such as what markets to open up, what product to offer or how much will the new product cost [22].

The explorative product development approach is an iterative process [22]. The primary market gets an early version of the new product. If you look at how often the product is sold, what has to be changed or what are the improvement proposals, then you rearrange the product and marketing approach according to what you learned and then try it again [22].

The User Centered Product Design focuses on the User Experience, but it does not contain a marketing or technological perspective [22]. This implies that potential users of a new product are part of the development process. The focus is on the discovery of the users’ needs, which are not precise defined at the moment. The goal is to develop, on the basis of user needs, non-existing products.

Then the users can get experience with the product previously and give feedback [22]. At this point it is important to explain the term ”Experience Design”. In Experience Design the human is central in the design [17]. Here the user is renowned as a ”expert”. The user knows best what he wants and what he needs and the designer plays the minor part. The Experience Design approach is certainly prescriptive driven. That means the designer can decide against a principle of the User Centered Design. Accordingly the handling of a product is deliberately limited or functions are missing. As a consequence a certain experience is generated [17].

4.1 Motives and Action

Creating a product concept in a user centered way, the designers hypothesize what are the needs and what does the user want to expire with the product in the future [22]. A. Mäkelä and J. Fulton Suri [27] describe a model of User Experience shown in figure 1. The

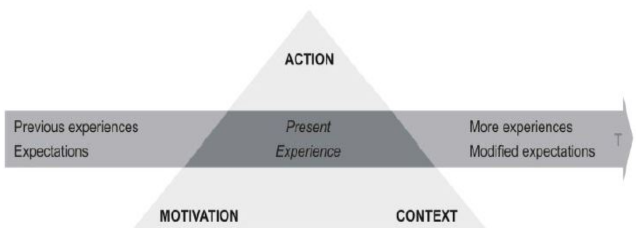


Fig. 1. Model of User Experience [27]

model explains three parts: the action, the motivation and the context [22]. On the time line the previous experiences and expectations of the user affect the present experiences and the present experiences evolve to more experiences and modified expectations. The context are for example people, places or things which are in their direct environment. A motivated action always takes place in a particular context. The motivation level describes the motivation to use a product and the action level explores how a user uses a current system [22].

The Self-Regulation Theory from C. Carver et al. [8] differentiate the levels of action between be-goals from do-goals to motor-goals. In the Activity Theory from V. Kaptelinin [23] the highest defined levels consist of activities and these fulfill specific motives. The activities themselves consist of goals and agreeing actions. The actions furthermore fulfill the goals [14] [23]. According to figure 2 M. Hassenzahl [14] illustrates the hierarchy of the goals and the related terms of the Activity Theory are listed in brackets. Therefore the be-goals describe why we use the product, the do goals

Fig. 2. Model of the three levels of goals [14]

describe what actions do we need to do to use the product or to reach a specific goal and at last the motor-goals, which describe the single operations [14]. An example would be to skype with a friend. The be-goal is that you miss for example a friend and you want to talk to him. The do-goals are to start the computer, start the application and then call the friend. Starting the computer for example operations are needed, which represent the motor-goals: you have to enter your username and your password for example.

4.2 Generate ideas among development team

To create new products it is important to understand the motives and actions of a user. The attention of the development team lies on the users, their needs, their physical and social environment and the situation to use the product [22]. There are many different ways to generate ideas among the development team, for example brainstorming. Choosing a technique depends on the team, but sometimes they consider new methods to generate new ideas. The Maypole Project Team [36] for example developed two idea generation techniques: the social map and the role-playing with toy characters. The team developed Maypole to discover and create a new product concept for children to communicate with the people of their social network [22].

The first technique is the social map. The map includes primary and secondary users and their places [22]. With the help of narrative examples the team gains an insight of the studied users and their needs in real-life situations. Afterwards the team is separated in pairs and every pair gets an image with the users and their places. Now the pairs have to think about new product concepts, which the user would use in the places in the image. The team has to think about the concept, but only with a limited time and the ideas have to be drawn on paper [22].

The second technique is role-playing with toy characters. The characters present the users and they are placed on a map of their environment. Based on the user research results the team members get different roles to play in a particular situation. While playing with the characters the team can think about new concepts or hit new problems [22].

Both techniques made the same sum of new ideas, but the ideas from the first technique were mainly wireless applications and the ideas from the second technique were mainly location-based [22]. A possible explanation would be that the second technique uses a single defined playboard for the team, whereas first technique uses different images.

4.3 Generate ideas with users

The second option to get new ideas is to generate them with the users. In this case it is possible to include the social context into the activity [22]. It is important to involve the consumer into the design process, because they are the people who will finally buy and use the product [1]. But every user is unique and not always comparable to others. K. Eason [11] describes that it is important to distinguish different types of users: the primary, the secondary and the tertiary type. The primary type is the kind of consumer who will utilize the product. The secondary type is the one who will occasionally apply the product and the tertiary one will be affected by the use of the product or decides to buy it [1]. It is also important to look at the needs an expectations of the people who will be affected by the product.

Like in the subsection before there are different techniques and methods to generate new ideas, for example [33]:

- background interviews and questionnaires: help to collect data based on the needs and expectations of users
- sequence of work interviews and questionnaires: help to collect data based on the sequence of work which is performed with the product
- focus groups: a group of stakeholders who talk about aspects, problems, results and requirements
- on-site observation: helps to collect data bases on the affected environment in which the product will be used [33]

The Maypole Project Team [36] used for example role-playing session. There is a group of users, who know each other and everyone gets their own role [22]. The group gets an example of a real world situation and then they have to continue playing. The role and the situation base on user research results. In the Maypole project there were additionally images on the walls of current technologies. But the users were not allowed to use them in their normal way, but they should invent new possibilities to use them in a new way to enhance the situation. In the center of idea generating in the area of User Experience are the users, their needs, their physical and social environment and the situation to use the product [22].

Regardless of whether the development team gets new ideas by generating them on their own or with the user, it is important to collect a lot of new ideas. The ideas can be compared to each other, can be combined or can be the foundation of new ideas. Therefore the first step in product development is the idea generating process. After you know which product you want to develop, you can go to the next step in product development: creating a prototype of the idea.

4.4 Implementation of an idea: The story

The base of prototypes are stories. Stories are a good way to share information. They are suitable for the User Experience Design process, because the system can be used in a real context [32]. It establishes a connection between the design ideas and the user, who will make use of the product, because it helps to let the user be at the center of the work. Therefore stories can help the developer to get a better understanding about the users and their experiences and help to create better products. Stories can be incorporated in the development process, because they characterize a context or a situation, they describe problems and show which new experience is needed and they can launch a design discussion to examine new concepts or a new design [32]. W. Quesenbery et al. [32] describe five parts of the UX process, where storytelling can be useful:

- gather input: collect stories from the users about their daily life or what they want
- examine user research and other information: clustering of stories and select the stories which are good examples with illustrating patterns and personas
- simulate or experiment with design ideas: the stories help to develop ideas and the emerging design has to base on real user needs
- test the design: test if the design works for the real user, the stories work well as test scenarios
- share or sell your ideas: stories can show why a design will work because they link the design and the inspiration of an idea

According to this, stories can be useful in many steps of the development process and storytelling is an important technique in the field of User Experience.

After creating a good story, a storyboard must be developed [32]. A storyboard shows the context and the events, because they can offer a wider view of the interaction. The storyboard can be compared with comics. There are visual images with a certain order and captions or figures who speak or think [32]. A storyboard furthermore is already a prototype [33]. Prototypes that base on stories allow the team to investigate a new vision [32]. The following section classifies the different kinds of prototypes and storytelling is assigned to one as well.

5 How to create Prototypes

With the help of prototypes designers have an effective way to examine the design ideas and the team can choose between alternatives [33]. There are many different ways why a prototype is useful: for example to try out the technical practicality of an idea or to test that a particular design direction is suitable with the rest of the product development [33]. The goal of User Experience Prototypes is that experience and knowledge are gained with the product or system. With the help of prototypes the development team gets feedback before developing a marketable product [22]. According to M. Buchenau and J. Fulton Suri [7] there are three different types of activities within the development and design process where Experience Prototyping is useful:

1. understanding current User Experience and context
2. studying and evaluating design ideas
3. telling ideas to an audience [7]

The first point explains that it is important to show the context and to identify aspects and design opportunities [7]. The questions that have to be answered are: What are the contextual, physical, temporal, social and cognitive factors to think about before starting the design? What is the key aspect of existing user experience? What are crucial factors that the design should retain? The goal is to get a high fidelity simulation of an existing experience which can not be experienced immediately, because the experience is for example unavailable, expensive or unsafe. The second point describes that it should be easy to explore possible solutions and the design team should be more informed about development of User Experience and the influencing components [7]. The focus is on particular artifacts, elements or functions. Prototypes of these elements and their interactive behavior can help to evaluate many ideas and to form User Experience. The focus of the last point is the user. The user has to understand the subjective value of a design idea. To get this understanding the user has to experience it directly. In this case the audience has to be persuaded [7].

Experience Prototyping is not about the development of a toolkit or a specific technique, it is about an attitude and a language to solve a design problem, but it is important to know that one single prototype is never enough [7].

5.1 Resolution and Fidelity

Prototypes which look finished imply that the design they present is near finalization [20]. It is possible that prototypes which look finished are made in the early design process (for example a concrete 3D model) and rough looking prototypes are made in the later design process (for example the prototype includes the overall structure but not all the visual details). To classify these different kinds of prototypes there are two terms: resolution and fidelity [20]. Resolution is according to S. Houde and C. Hill [20] interpreted as the amount of detail and fidelity is the closeness to the eventual design. Resolution is differentiated in high and low resolution. High resolution means in view of the definition of Houde and Hill that a prototype shows what and how something will work [20]. Therefore it is not important how something is implemented as long as it is implemented. For example if the user can set a value on a slider, this can be implemented with a balloon filled with helium. It can be used to show the value by letting it fly higher or lower. Whereas a low resolution prototype does not have the amount of detail, there would not be the possibility to set a value.

Fidelity prototypes can be differentiated in low, high and mixed fidelity prototypes. According to J. Hocko [19] the choice which degree of fidelity has to be chosen to implement depends on many factors:

- a detail evaluation of the consisting research and statements
- the culture of an organization and the political structure
- actual software development levels and approaches
- the purposes the prototype has to serve
- the ability and flexibility of each project team member

Which resolution or fidelity is needed depends also on the nature of the audience [19]. An interactive storyboard for example is useful for the design team but is not comfortable for the supporting organization [20]. There can be more other estimates to decide which degree of fidelity is suitable, because each project team is different [19]. It is important to say that the degree of visual or behavioral elaboration of the prototype is not compulsory the progress of the design or a certain stage in the process [20].

5.1.1 Low Fidelity Prototypes

This kind of prototype does not look like the final product [33]. Low fidelity prototypes are mainly paper based. It is realized with a pen and a paper or something like a whiteboard, so with low-cost methods. The developer can easy implement such kinds of prototypes because it is low level [2], cheap and quick to produce, so they support the exploring of alternate design concepts and ideas. That means these prototypes are important in the early stages of the development process [33]. The problem of this prototypes is that it is difficult for the user to feel the system or to get experience with it, because he cannot play with it as he would do it in reality [2]. Another problem is the limited error proofing or the navigation limitations [33].

An example for a low fidelity prototype is the storyboard mentioned before. It shows the single steps which a user perhaps goes through when he uses the product [33]. Another typical example would be a paperprototype [2]. Figure 3 shows a paperprototype of an Event Manager. The user can see the interface and can for example complain about buttons he can not understand. The user can play with that prototype if there is a storyboard which describes the single functions of the product. But the user can not get for example the time it would take to perform an action, he does not know how the response would feel like or he can not recognize if there are unexpected actions [2].

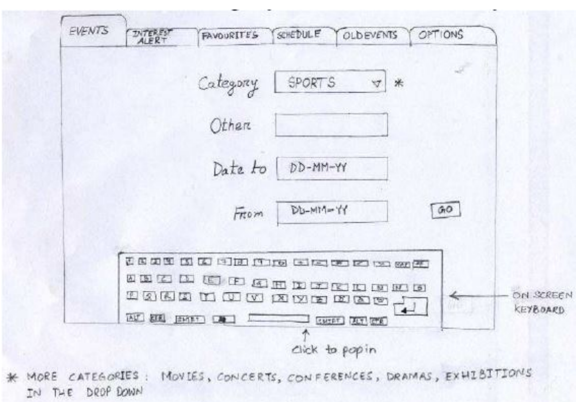


Fig. 3. low fidelity prototype of an Event Manger [2]

5.1.2 High Fidelity Prototypes

The high fidelity prototype is considered as an advanced prototype [2]. They are associated with more costs, because they use some tools, languages and various high tech equipment [2]. These kinds of prototypes can be tested in an environment familiar to the user [22]. The user can get a better understanding, how the product will look and feel and how it will work. He can get an image about the efficiency of the product and the input responses [2]. Therefore high fidelity prototypes are completely working, interactive, user-driven and the user gets a look and feel of the final product [33]. These kinds of prototypes are suitable for Experience Prototyping but are limited by arising costs and the long time they need to be created [2] [33].

An example is the Clique Trip project created by M. Knobel et al. [25] from the BMW Group Research and Development. Clique Trip is an in-car system. If there is a group of friends traveling in different cars, the system can show the distance between the cars and if the cars are close enough the friends can chat with each other. The final prototype was implemented with a smart phone using HTML, Javascript, PHP, AJAX and MySQL. The application was integrated in the infotainment system written in Adobe Flash and ActionScript 3. The application was implemented in the car and was testes on the road [25]. Figure 4 shows one view of the display, what the user can see while driving. It shows that the second car is behind the first car with a large distance. Therefore the prototype is completely working,

Fig. 4. Clique Trip project: one view of the display. It shows a large distance between the cars, the other car is behind [25]

interactive, user-driven and the user gets a look and feel of the final product [33]. Therefore this prototype is classified as a high fidelity prototype.

5.1.3 Mixed Fidelity Prototypes

Mixed fidelity prototypes are a mix of low fidelity and high fidelity prototypes [2]. The prototype is cost efficient. Only the parts which have to be high fidelity causes the costs. The other parts which do not need to be high fidelity can base on paper for example which does not affect the costs [29]. This kind of prototyping is useful for Experience Prototyping, because the users still have a good understanding of the product and it is cost efficient [2].

In the Maypole project [36] for example the users tested one of the concepts in a laboratory environment. They got blank models of the product and have then explained what their first impressions are and how the product looks and feels. After that they got user scenarios in the form of storyboards and had to think about other potential scenarios which would be possible in their own life. Next the users got prototypes, which include the main interaction style of the product concept and had to give feedback about it. The last test was a discussion. The users had to describe what they think for whom the product concept would be appropriate and what the development team has to change about the product to make it more reasonable [22]. In this tests many kinds of prototypes are deployed. There are low fidelity prototypes for example the scenarios in form of a storyboard, there are mixed fidelity prototypes which present the look and feel and at least there are high fidelity prototypes which include the main functions of the product. The Maypole project is an good example that one single kind of prototype is not enough. It is important to use for example different fidelity prototypes which produce diverse results with the user, because they pursue different goals in the development process. The resulted issues improve this process and new prototypes can be developed.

5.2 The Model

To identify which kind of prototype is the best choice to use in Experience Prototyping there is a model. The model is developed from S. Houde and C. Hill [20] and shows a three-dimensional space. The dimensions represent significant aspects of the design of an interactive product or system (figure 5):

- The role: What is the function of the product? Why is it useful for the user's life?

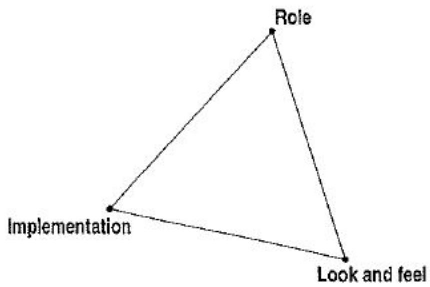


Fig. 5. ”What prototypes prototype”, the model [20]

- The look and feel: What is the experience of using this product? Looking at the concrete sensory how it looks like, feels like and sounds when the user uses it.
- The implementation: Which techniques and components are used to make the product or system work?

The model is drawn isosceles and oblique which means that no aspect is more important than another aspect [20]. The goal of the model is that if there is a design problem the model divides the different design aspects into three categories of questions. These questions require different approaches to prototyping. The implementation normally needs a working system to be built, the look and feel aspect needs specific user experience to be emulated or created and at least the role aspect needs the context, where the product would be used, to be introduced. If you know which design question you must answer it helps to decide which prototype to use [20].

An example for a look and feel prototype is the Architect's Computer developed by the Apple Design Project from Apple Computer Inc [20]. The team wanted to design a portable computer for architects. They wanted to know how an architect would carry the computer, what else he takes with him and which tasks he had to do during visiting a building. For this the design team uses a pizza box and gave it the weight of the computer. With the help of this low fidelity prototype the architect can get a look and feel about the product if he would take it with him around [20]. The focus of the Apple Design Project is to find out what is the daily routine of a user. According to this they have to know how a user handles the product and what the experiences are. The design problem in this case can be solved by a look and feel prototype. The development process is in the early phase and therefore the best kind of prototype would be a low fidelity prototype. The prototype has to have the characteristic to let the user have experience with it.

6 UX Evaluation Methods

In the evaluation process User Experience should have the focus on the rising experiences and their attributes and not on the products and their attributes [13]. The evaluation context focuses on the measure of satisfaction of needs. Thereby it is helpful to use a model that combines positive emotions with the satisfaction of universal needs and clarifies the coherence between needs and specific attributes [17]. The developed prototypes have to be evaluated to find problems or improvement proposals. In UX Evaluation methods the focus is on lived experiences [24]. According to Vermeeren et al. [37] there are about 96 different methods for User Experience evaluation [1]. With such a large number of methods it is difficult to determine, which ones are the best. In this section three different methods are picked out and are explained: the PANAS questionnaire, the AttrakDiff questionnaire and the psychological needs.

PANAS means Positive Affect Negative Affect Schedule [39]. The Positive Affect mirrors the scale to which the user feels enthusiastic, attentive and active. Unlike the Negative Affect, this dimension includes an amount of uncomfortable mood states, like anger, fear or nervousness. With the help of a questionnaire it can be analyzed, which mood of the user is affected. A large number of research areas developed and studied a lot of PA and NA scales [39]. The mood questionnaire was developed by Zevon and Tellegen (1982) [40]. It includes 60 descriptors organized in diverse orders on a single page. To have reliability of the PANAS scale it is suggested to have ten terms of the PA and the NA. That means you have to identify which PA and NA terms you need to evaluate your project [39]. The general dimension scales are for PA [38] according to D. Watson [39]:

> ”attentive, interested, alert, excited, enthusiastic, inspired, proud, determined, strong and active” [39].

and for NA according to D. Watson [39]:

> ”distressed, upset, hostile, irritable, scared, afraid, ashamed, guilty, nervous, jittery” [39].

The different points of the scale are named according to D. Watson [39]:

> ”slightly or not at all, a little, moderately, quite a bit and very much” [39].

There are also eight different temporal information, which the user has to rate how he feels at the moment, today, during the past few days, the past week, the past few weeks, the past month, the past year and how he feels in general [38].

The AttrakDiff evaluates the feelings of a user about the product [16]. It is a questionnaire to measure the perceived hedonic and pragmatic quality. The AttrakDiff 2 consists of a self constructed questionnaire in form of a semantic differential. There are 28 seven stage items and every item has contrary adjectives on their endpoints, for example good and bad [15] [16]. The average of the items forms the values of pragmatic quality, hedonic quality and attractiveness [15]. The different terms are explained according to M. Hassenzahl et al. [16]. Pragmatic quality describes the recognized abilities of a product to reach a goal because of its useable functions. The hedonic quality is divided into stimulation and identity. Stimulation is the ability of a product to satisfy the need for improvement of own knowledge and skills. Whereas identity means the ability of a product to communicate self-worth messages with others. The attractiveness at least describes the global positive and negative assessment of a product [16].

K. Sheldon et al. [35] describe the psychological needs in their paper and which ones are fundamental for humans. They limit ten different items from different theories of researchers. The items are:

> ”autonomy, competence, relatedness, self actualization-meaning, psychical thriving, pleasure-stimulation, money-luxury, security, self-esteem, popularity-influence” [35].

There are three different studies in this paper. The first identifies which items are present when people try to satisfy their experiences. The second one identifies which qualities of experience show best the change in positive and negative effect according to a given event. In the last study the results from the other two studies are proven regarding to the replicability to a longer time frame. The results from these studies are that the characteristics that make people happier and are qualified as psychological needs are according to Sheldon et al. [35] autonomy, competence, relatedness and self-esteem. Another item is security, but this one becomes noticeable when there is deficiency. Less important are pleasure-stimulation, self-actualization meaning, popularity-influence and physical thriving. Sheldon et al. would reject their need status. Whereas money-luxury should also get the need status [35].
It is possible to combine the statistic of the needs and the PANAS

[1] http://www.allaboutux.org/all-methods

scale to identify whether the need is associated positive or negative. Whereas the AttrackDiff is independent from PANAS and the need statistic because the focus here is on the aesthetic and attractiveness. To link a connection between attributes and a concept is not a new technique [17], there is a qualitative evaluation method called laddering. Laddering bases on a repertory grid technique and is a structured questioning technique. The different attributes are organized hierarchically [9].

7 Discussion

After the explanation of the development process, which starts with the idea generating followed by the developing of prototypes and evaluation, it is necessary to point out the differences, similarities and the coherence between User Experience and Usability. As described earlier UX is part of the User Centered Product Design with the focus on the users' needs. Usability may be a part of this approach for example in form of Usability testing to test if the user can deal with the product. If he is not able to use it in the right way, it may be difficult to achieve the right experience. The conventional continuous product development and the explorative product development focus on Usability. Both approaches are iterative processes and require techniques where Usability is necessary.

Usability and UX have the same idea generation methods for instance brainstorming. However there are also specific techniques which are primary used in Usability or User Experience. The UX techniques base on the experience and emotion e.g. role playing. It is important to get a better comprehension about the feelings of the user. Usability mainly tries to develop new products, where the focus is on the performance of a product and that the user can use the product intuitively. The storytelling is an important technique in UX as well as in Usability. Scenarios play an important role in the user tests. In Usability the scenarios function as an example whereas in UX there should be a connection between the user and the story. It would be helpful when the user already was in such a situation or he empathizes with the story because then he is able to set up emotions with the scenario.

Prototypes play an important role in every product development process. According to the model from Houde and Hill the UX prototypes are placed in the look and feel corner. Whereas the Usability prototypes focus on the implementation which is important to make the product work in the right way. It is important to reduce the response time of the system and speed up the system navigation. The role character of a product is significant for both Usability and User Experience. The product should play an important role in the user's life whether to have experience with it or to facilitate the user's daily life. The focus of Usability for example lies on the task performance, but UX focuses on lived experiences. For UX it is not important to count the numbers of clicks [5], it is more important to detect what the user feels and thinks when he uses a product or systems [24].

In the evaluation process the relationship between User Experience and Usability methods are intertwined. A. Vermeeren et al. [37] consider that Usability is subsumed by User Experience. UX evaluation comprises the existing methods for Usability evaluation. According to N. Bevan [5] there are also differences between Usability and User Experience measures. The difference of the focus between task performance and enjoyment leads to different causes during the development. According to N. Boven [5] Usability focuses in the field of design and evaluation on:

- overall efficiency and effectiveness and the user's comfort and satisfaction
- the product should be easy to use and the evaluation of the product attains identifying usability problems
- perhaps learnability

Whereas User Experience deals with:

- How do people interact with the product over a specific time period, what are they doing and why?
- the performance of the hedonic goals of stimulation, identification and evocation have to be maximized and emotional responses have to be related [5]

In summary Usability and User Experience are different approaches, but both play an important role in the field of Human Computer Interaction and are significant in the development process in their own field of application.

8 Summary

This paper gives answers to five questions. The first is what is the meaning of Usability, User Experience and experiences? Usability was defined according to the ISO standard, whereas User Experience has several definitions. The definitions focus on the user's interaction and the resulting factors. Experiences are in this case an important factor and are classified in four threads: the sensual, the emotional, the compositional and the spatio temporal thread [28]. Experiences generate a relationship between the user and the product.

The second question is what is to do before developing prototypes? There are two possibilities to generate new ideas, among the development team and with the user. There are several techniques like brainstorming, role-playing, the social map or questionnaires. The questions why the user uses a product, what the actions are and how the user performs the single actions influence the process of generating ideas [14]. The storytelling is a useful technique. It can be used in many different ways and is the base for storyboards and prototypes.

The third question is how to create prototypes? Experience Prototyping is useful to understand current User Experience and the context, to study or evaluate design ideas and to tell the ideas to the audience [7]. The terms resolution and fidelity describe the amount of detail and the closeness to the eventual design [20]. Prototypes can be distinguished in high and low resolution and in low, high and mixed fidelity prototypes. A low fidelity prototype is for example a paperprototype. A high fidelity prototype completely works, is interactive, user-driven and the user obtains a look and feel of the final product [33]. The mixed fidelity prototype combines the approaches of low and high fidelity prototypes. With the help of the model from Houde and Hill [20] the different design aspects can be divided in three categories of questions and these questions require different approaches of prototyping. The three dimensions of the model are the role, the implementation and the look and feel.

The next question is what are the UX Evaluation Methods? According to Vermeeren et al. [37] there are about 96 different methods for User Experience evaluation. In this paper the PANAS questionnaire, the AttrakDiff questionnaire and the psychological needs were discussed. The last question is what are the differences and similarities between Usability and UX in view of the development process? The main difference is that Usability deals primary with the user's cognition and performance in interactions, whereas User Experience handles with non utilitarian aspects of interactions [26]. This focus affects the whole development process and influences the idea generating, prototyping and the evaluation of the process.

User Experience and Experience Prototyping can help to create new innovative products. The needs, experiences and emotions play a necessary role and promote new ideas when the user is in the center of the development process. In further research there should be a consistent definition [26] of User Experience which would facilitate the understanding of the term and there would be no longer a danger of confusion with Usability.

References

[1] C. Abras, D. Maloney-Krichmar, and J. Preece. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, 37(4):445–56, 2004.

[2] Y. Ansar-Ul-Haque. Enhancing experience prototyping by the help of mixed-fidelity prototypes. In *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology*, pages 468–473. Association for Computing Machinery, 2007.

[3] N. Bevan. Quality and usability: a new framework. *Achieving software product quality, van Veenendaal, E, and McMullan, J (eds)*, 1997.

[4] N. Bevan. Ux, usability and iso standards. *Conference on Human Factors in Computing Systems 2008*, pages 1–5, 2008.

[5] N. Bevan. What is the difference between the purpose of usability and user experience evaluation methods. In *Proceedings of the Workshop User Experience Evaluation Methods*, volume 9, 2009.

[6] L. V. Boven and T. Gilovich. To do or to have? that is the question. *Journal of Personality and Social Psychology*, 85(6):1193–1202, 2003.

[7] M. Buchenau and J. F. Suri. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 424–433. Association for Computing Machinery, 2000.

[8] C. S. Carver and M. F. Scheier. *On the self-regulation of behavior*, volume 441. Cambridge University Press, 2001.

[9] C. Corbridge, G. Rugg, N. Major, N. Shadbolt, and A. Burton. Laddering: technique and tool use in knowledge acquisition. *Knowledge Acquisition*, 6(3):315–341, 1994.

[10] DIS and ISO. 9241-210: 2008. *Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems*.

[11] K. D. Eason. *Information technology and organisational change*. CRC Press, 2005.

[12] M. Glanznig. User experience research: Modelling and describing the subjective. *Interdisciplinary Description of Complex Systems*, 10(3):235–247, 2012.

[13] M. Hassenzahl. User experience (ux): towards an experiential perspective on product quality. In *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine*, pages 11–15. Association for Computing Machinery, 2008.

[14] M. Hassenzahl. *Experience design: Technology for all the right reasons*, volume 3. Morgan & Claypool Publishers, 2010.

[15] M. Hassenzahl, M. Burmester, and F. Koller. Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. *Mensch und Computer 2003. Interaktion in Bewegung*, pages 187–196, 2003.

[16] M. Hassenzahl, M. Burmester, F. Koller, H. Brau, S. Diefenbach, M. Hassenzahl, F. Koller, M. Peissner, and K. Röse. Der user experience (ux) auf der spur: Zum einsatz von www. attrakdiff. de. *Usability Professionals 2008*, pages 78–82, 2008.

[17] M. Hassenzahl, K. Eckoldt, and M. T. Thielsch. User experience und experience design - konzepte und herausforderungen. pages 233–237.

[18] M. Hassenzahl and N. Tractinsky. User experience-a research agenda. *Behaviour & Information Technology*, 25(2):91–97, 2006.

[19] J. M. Hocko. A practical evaluation of the prototype fidelity debate.

[20] S. Houde and C. Hill. What do prototyps prototype? In *Handbook of Human-Computer Interaction*, pages 367–381. Elsevier Science B.V., 1997.

[21] ISO and WD. 9241-11. ergonomic requirements for office work with visual display terminals (vdts). *The international organization for standardization*, 1998.

[22] A. Kankainen. *Thinking model and tools for understanding user experience related to information appliance product concepts*. 2002.

[23] V. Kaptelinin. *Acting with technology: Activity theory and interaction design*. Mit Press, 2006.

[24] J. Kaye. Evaluating experience-focused hci. In *Conference on Human Factors in Computing Systems 2007 extended abstracts on Human factors in computing systems*, pages 1661–1664. Association for Computing Machinery, 2007.

[25] M. Knobel, M. Hassenzahl, M. Lamara, T. Sattler, J. Schumann, K. Eckoldt, and A. Butz. Clique trip: Feeling related in different cars. In *Proceedings of the Designing Interactive Systems Conference*, pages 29–37. Association for Computing Machinery, 2012.

[26] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort. Understanding, scoping and defining user experience: A survey approach. *Conference on Human Factors in Computing Systems 2009 User Experience*, pages 719–728, 2009.

[27] A. Mäkelä and J. Fulton Suri. Supporting users creativity: Design to induce pleasurable experiences. In *Proceedings of the International Conference on Affective Human Factors Design*, pages 387–394. Asean Academic Press, London, 2001.

[28] J. McCarthy and P. Wright. *Technology as Experience*. MIT Press, 2004.

[29] M. McCurdy, C. Connors, G. Pyrzak, B. Kanefsky, and A. Vera. Breaking the fidelity barrier: an examination of our current characterization of prototypes and an example of a mixed-fidelity success. In *Proceedings of the Special Interest Group on Computer-Human Interaction conference on Human Factors in computing systems*, pages 1233–1242. Association for Computing Machinery, 2006.

[30] J. Nielsen and D. Norman. The definition of user experience. `http://www.nngroup.com/articles/definition-user-experience/`. visited 03.05.2014.

[31] T. Nurminen. The meaning of user experience. `http://www.userintelligence.com/ideas/blog/2011/04/meaning-user-experience`, 2011. visited 17.05.2014.

[32] W. Quesenbery and K. Brooks. *Storytelling For User Experience - Crafting Stories For Better Design*. Louis Rosenfeld, USA, 2010.

[33] Y. Rogers, H. Sharp, and J. Preece. *Interaction design: beyond human-computer interaction*. John Wiley & Sons, 2011.

[34] V. Roto, E. Law, A. Vermeeren, and J. Hoonhout. User experience white paper. *Bringing clarity to the concept of user experience*, 2011.

[35] K. M. Sheldon, A. J. Elliot, Y. Kim, and T. Kasser. What is satisfying about sytisfying events? testing 10 candidate psychological needs. *Journal of Personality and Social Psychology*, 80(2):325–339, 2001.

[36] M. P. Team. Maypole highlights: what makes kids tick? *interactions - Special issue: the digital hug*, 6:80–83, 1999.

[37] A. P. Vermeeren, E. L.-C. Law, V. Roto, M. Obrist, J. Hoonhout, and K. Vnnen-Vainio-Mattila. User experience evaluation methods: Current state and development needs. *Nordic Conference on Human-Computer Interaction 2010*, pages 521–530, 2010.

[38] D. Watson and L. A. Clark. The paanas-x: Manual for the positive and negative affect schedule - expanded form. *Department of Psychology Publications*, pages 1–24, 1999.

[39] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.

[40] M. A. Zevon and A. Tellegen. The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology*, 43(1):111, 1982.

Indirect Touch Interaction

Werner Eckert

Abstract— Direct touch interaction becomes more and more common in the last years with the introduction of smart-phones and tablets. But in regular working environments the drawbacks still prevent the adoption of touch input and should be examined. Indirect touch interaction can be a solution for future desktop environments to face this disadvantages.
This paper investigates the current work indirect touch interaction and tries to pin down methods and concepts to introduce it to classic desktop environments. For this it looks through Gilliot's work about device size, peripheral field of view, scale effects and aspect ratio in the first part to show advantages and drawbacks in user performance. In the next step concepts to re-introduce the three state model for pointing tasks are examined. Current techniques for this are mostly pressure or gesture based and do perform differently. After all, people use normally both hands to interact with the environment. So in the last part the asymmetric bi-manual concept of workspaces pinned to the left hand are introduced which have advantages especially when using very big displays like a power-wall. At the end some techniques for symmetric interaction are summed up which haven't been investigated for indirect touch interaction, but can extend the input space.

Index Terms—indirect touch input, indirect touch interaction, three-state-model, three-state-model implementation, peripheral vision, bi-manual touch interaction

1 Introduction

In the last years smart-phones and tablets have become an every day device and so on brought touch sensitive displays to a broad audience. By now, many studies have shown the advantages. So touching directly an element for manipulation is perceived more naturally by users than handling a mouse [6] and especially novices benefit from a steeper learning curve and the fact that there are no moving parts like keys. Thus, the technology is particularly appropriate for mobile devices as well as in public settings such as ticket vending machines.

On the other side touch-sensitive surfaces still haven't replaced keyboards and mice in production environments and so, the benefits of touch input can not be used. That is hardly surprising since direct touch-sensitive input has like any other technology also its disadvantages. The fat-finger problem addresses the fact that the finger just doesn't touch the surface point-like but does as a large area. This reduces the precision and increases the error rate [1], but that contradicts the demands of modern work places. Moreover, hands and arms cover elements on the screen with which the user wants to interact. This is especially a drawback when investigating large data-sets. And last but not least the limited arm length prevents interacting with distant or very large screens by direct contact. This also limits the use of big screens and moreover of installations like the power-wall. Owing to all these drawbacks it becomes clear why desktop environments still haven't adapted touch interaction. The separation of the input surface from the monitor like illustrated in figure 1 can overcome these disadvantages and can introduce the benefits of touch interaction to classic desktop environments.

Indirect touch interaction is topic of current investigations and pros and contras have been identified. Especially during long operations on classic desktop computers indirect touch interaction can retain the benefits of touch-sensitive surfaces and smooth out disadvantages. There is no danger that the user obscures the display output by his hand or arm position then using indirect interaction. It also makes sense from an ergonomic standpoint because the user just naturally sits up and looks straight [10]. Less tension and muscle pain in the neck are the result [5, 12]. Furthermore with indirect touch interaction working with both hands / arms on a horizontal plate

Fig. 1. Possible working-place with indirect touch interaction [13]

can be preserved, which is less strenuous and even faster [3, 9] and allows many possibilities for designing a rich design space, because all fingers of both hands can be used to perform input.

On the downside, the advantage of direct touch, that one can directly interact with the object, is reduced in indirect touch settings. Moreover, the users' accuracy reduces because he does not know the current position when not touching the surface. Finally it may be confusing and have drawbacks when the touch surface has different dimension than the monitor [7].

This paper investigates the possibilities to overcome drawbacks of indirect touch interaction to enrich classic desktop environments with touch interaction. There are many parameters for the basic setting of the work place, so in the first part the connection between surface size, monitor size and aspect ration is examined. Considerations for designing are shown in extra. Secondly, methods to re-introduce the tracking-state are showed, so the user is better informed about the current cursor position. Knowing that he can move for example the current selected tool to a position more preciser before engaging it. Finally this paper explores techniques to use the rich possibilities in designing the design space and how the left hand can support the right hand in a natural way.

2 Exploring Indirect Touch Settings for Work Places

While direct touch interaction put itself forward for mobile input devices, indirect touch interaction orientates itself towards classic work environments where input and output devices differ. Indeed keyboard

- *Werner Eckert is studying Media Informatics at the University of Munich, Germany, E-mail: werner.eckert@campus.lmu.de*
- *This research paper was written for the Media Informatics Advanced Seminar, 2014*

and mouse featured properties can be found also in indirect touch input and must be re-examined under this new context. The following features are identified for this work:

- Device size and peripheral field of view
- Scale effects and aspect ratio
- The hover state for touch
- Bimanual interaction techniques

2.1 Principal workplace considerations

In classical desktop setups the mouse does not define a limited input surface. In contrast to that we need a concrete area for tracking the fingers. Gilliot et al. [7] investigated how the dimensions of the tracking area are related to the output and if seeing the touch surface improves the interaction.

2.1.1 Don't loose track of your finger

In the first experiment Gilliot et al. wonder if having the input surface within field of view improves user performance and which surface area is easiest to reach. The variables in this discrete 2D pointing task were the input device size, the input condition, the target position and the target size. The input device came in two sizes. An iPad 1 offered a 196x147 mm large surface and an iPod Touch 3 offered a 66x50mm surface. The input condition were one handed, one handed with blinders and two handed with blinders. Figure 2 shows the cover for the different touch input surface sizes and the glasses with the blenders. The blinder was fixt to glasses so it blocked the peripheral view of the

Fig. 2. Cardboard overlay to simulate different input touch areas on the iPod Touch (left) and paper blinders taped to plastic glasses to prevent peripheral view (right) [7].

hands *(see figure 2 and 3)*. The targets had three widths sizes (10mm, 20mm and 40mm) and were distributed at best over the screen. The participants had to run the test several times always with a change in one parameter.

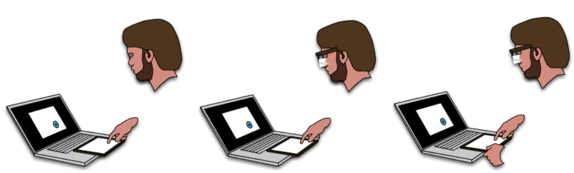


Fig. 3. Different settings for uni-manual (left), uni-manual with blinder glasses (middle) and bi-manual with blinder glasses (right) interaction [7].

Figure 4 shows the number of failed attempts across the different variables. One finding is that the target error was slightly less for targets in the center and in the north-west sector. Surprisingly larger targets had a higher number of failed attempts for being hit. Gilliot et al. suspect that the users may thought that larger targets are easier to hit and thus did care less. But above all Gilliot et al. [7] showed, that the targeting error is differed between the distinct input device sizes. Smaller devices outperform larger ones in targeting errors. Moreover, he et al. concluded that having the input surface within your field of view supports hitting a target while forming a reference frame with

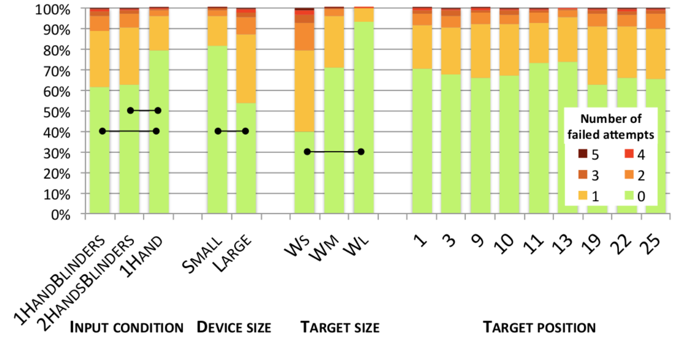


Fig. 4. "Number of failed attempts across INPUT CONDITION , DEVICE SIZE , TARGET SIZE and TARGET POSITION (from left to right). Connections between bars represent statistically significant differences" [7].

your non dominant hand does not provide any benefits.

Based on the experiment Galliot et al. recommend two approaches for designers [7]. First of all the investigation shows that being able to look at the input surface improves the performances. So users should see the touch input surface in absolute indirect-touch pointing tasks and be able to distinguish between it and the display. Moreover according to the outcomes targets near the center or the corners are easier to select. For right-handed users elements near the right-top corner are general easier to hit. So a second recommendation is, that when creating a graphical user interface you should consider the handedness of the user and thus put important elements in the middle or the north-west quadrant of the workspace.

2.1.2 Scale effects and aspect ratio

In an second experiment Galliot et al. [7] examined the impact of forms factors of the display on absolute indirect touch pointing performance. Again the participates had to do a 2D targeting test with the difference that the output display was now a 50" monitor in front of them. Part of the monitor was covered by different blenders to simulate different display sizes(H_S = 74mm, H_M = 147 mm, H_L = 294mm) and aspect ratios (R_M = 4:3, R_L=16:9, R_{XL} = 32:10). The input condition was always unimanual without blender glasses. Like shown in figure 5 a small display hight affects the number of failed attempts negatively while the medium aspect ratio, as used on common displays, performed worst. Although larger sized targets thought to be hit better than smaller ones the smaller one were hit better. Galliot et al. [7] presume that users care less about precision when hitting larger targets which appear being easier to hit.

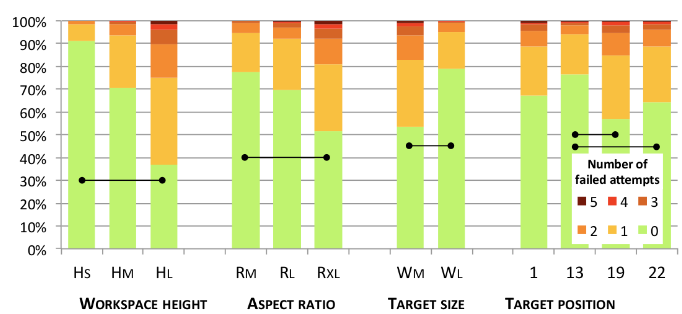


Fig. 5. "Number of failed attempts across WORKSPACE HEIGHT , ASPECTRATIO , TARGET SIZE and TARGET POSITION (from left to right). Connections between bars represent statistically significant differences" [7].

As reported by the experiment the success rate and the targeting error are independent from the display size but are not from the aspect ratio. So Galliot et al. [7] recommend similar rations for input surface and the monitor to boost the success rate and to decrease the targeting error.

Also they showed that participants were more precise with smaller targets which seems counter-intuitive. They determined that the limit to the target size was about the width of the finger tip. According to this reliably interaction with objects can only be obtained when considering the minimum target size in motor space for absolute indirect-touch interfaces.

2.2 Bringing in hover state for touch

Comparing the mouse as an input device to the input via touch, it is noticeable that the touch input still does not offer a state model like seen in the figure 6. It just offers the two states out-of-range (hand is off the surface) and engaged (finger presses to the surface). But especially with indirect touch input the tracking state may assist the user by providing additional information about the current position. Voelker et al. [13] show in their work different techniques to implement such a state model *(see figure 6)* for indirect touch interaction.

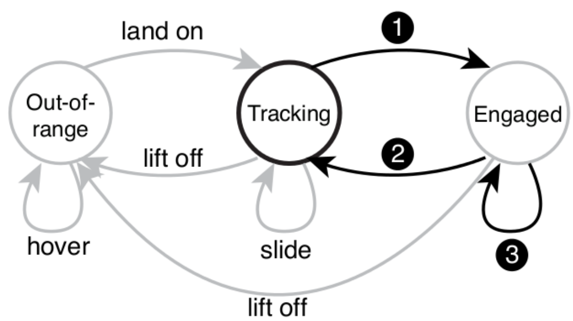


Fig. 6. Three-state interaction model [4].

But moreover, Voelker et al. postulate four design considerations which their methods should meet. First the methods should be applicable to individual fingers. This ensures that the technique retain the expressiveness of multi-touch input, can be used without identifying the single fingers and can be used by several users. Second it should be very hard to unintentionally switch between the states and last but not least the practice over a longer time should not stress the hand. Moreover the awareness of the current state should also be one of the design considerations in my opinion.

In the work of Voelker et al. all techniques assume that if the user doesn't touch the surface the out-of-range state is engaged and if the user does touch slightly the surface it is switched to the tracking state.

2.2.1 Pressure based techniques

The first group of techniques are pressure-based. This means they work with the physical pressure of the finger on the surface. However Pawluk and Howe have proven that the pressure strength is directly related to the contact area of the finger tip [11], and therefore these techniques are also applicable to touch-sensitive surfaces without pressure sensors. Nevertheless keeping up the pressure level, especially while moving, could be uncomfortable because of the frictions between the skin and the surface [13]. Moreover, every user has different finger sizes and can build up pressure differently.

It has to be mentioned that pressure based techniques will not suit mobile touch surfaces because the user has to build up additional pressure from the other side to hold the device in place.
Voelker et al. [13] discuss these two pressure based techniques:

- pressure quasimode

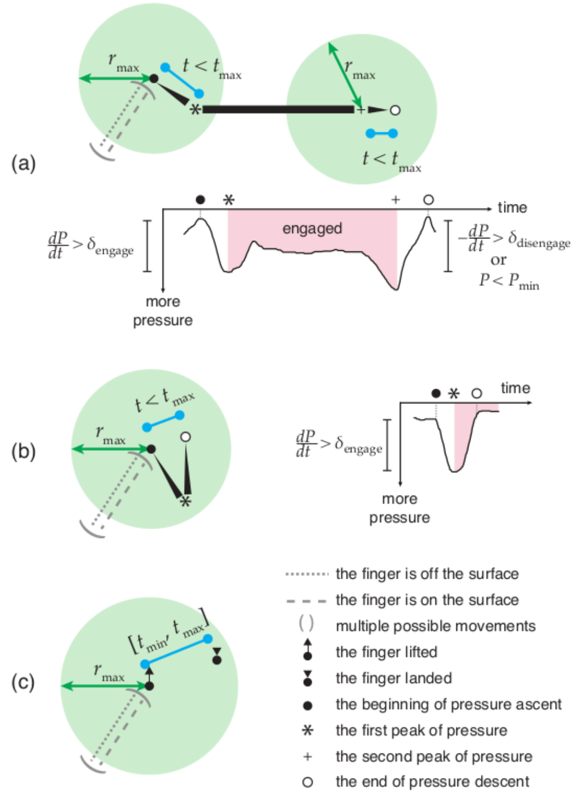


Fig. 7. pressure quasimode (a), pressure switch (b), lift-and-tap (c) [13].

With this method the state change is triggered by fast building up a pressure level with the finger within a certain time like illustrated in the graph of figure 7a. Keeping up the pressure level will hold the engaged state. To return to the tracking state the pressure of the finger is decreased really fast or under a minimum pressure threshold.

Although the user is always aware of the current state, keeping up a high pressure level can be exhaustive for the muscles and repetition may lead to injuries with this method [9]. To my mind this is a big drawback especially for weak or challenged people.

- pressure switch

  Here, a short, intense pressure by the finger toggles from the engaged state to the tracking state and back *(seen figure 7b)*.

  An advantage of this method is that the danger of straining the muscles were taken out. Moreover, he may forget in which state he currently is.

2.2.2 Gesture based techniques

Gesture based techniques avoid the need of pressure sensors and so can be applied to every touch input surface. Voelker et al. [13] discuss these two gesture based techniques:

- lift-and-tap

  Lift-and-tap is a technique in which a very short lift of the finger from the surface triggers the engaged state as seen in figure 7c. Important is that the finger has to touch the surface at the same spot within some margin. Then the engaged state stays until the finger is lifted out of range.

While tapping with the index finger is easy, this task may become challenging on multi touch input surfaces due to the limited freedom of movement of the other fingers. Maybe this can be overcome when you apply the three-state model only for the index fingers and a two-state-model to the rest of the fingers. This can be topic of future research.

- hold

  Here, the idea is not to move the finger for a certain time. After this it is switched to the engaged state. Lifting off the finger switched the state to the out-of-range state.

  This technique is also not suitable for multi touch interaction due to the possibility that the user unintentionally rests other fingers on the surface. Moreover, always restarting in the out-of-range state may slow down the working speed like always have to leave the mouse after every click.

2.2.3 SimPress Clicking

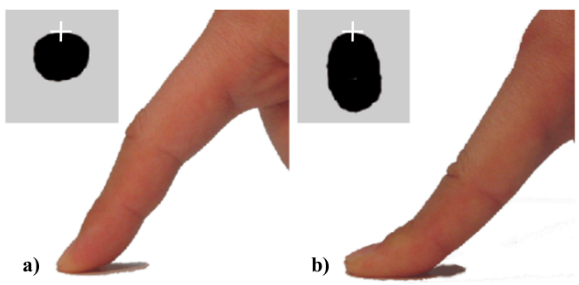


Fig. 8. "A small rocking motion of the users finger triggers the SimPress clicking technique: a) tracking (hover) state, b) dragging (click) state. (The top left corners show the actual area of contact detected by our device as well as the stabilized cursor location.)" [2]

Current touch-sensitive displays doesn't have pressure sensors nor can they track the finger without touch. So Benko et al. [2] suggested the SimPress (**Sim**ulated **Press**ure) Clicking in their paper. "SimPress requires the user to apply a small rocking motion with their finger in order to perform a "click", as seen in figure 8 [2]. In addition Benko et al. fixed "the cursor to the top-middle point [of the finger tip, so] the user is also able to make a more drastic change in the contact area without significantly disturbing the cursor location which aids in reduction of the unintentional clicks"[2]. Two thresholds suppress errors through noise or hand tremor.

Although Benko et al. examined the SimPress Clicking, he did it just for direct input surfaces. But in my opinion the SimPress Clicking outperforms the pressure based and also the gesture based techniques because they either need to use high pressure or you are in risk to forget the current state. So this technique should be investigated for indirect touch interaction in the future.

2.3 Use both of your hands

Current workplaces offer a keyboard and a mouse to interact with the computer. Especially at professional work places users use their left hand to use short-cuts on the keyboard to support the right hand which performs accurate actions. To capture the full potential of both hands for indirect touch displays Malik et al. [8] developed several uni-manual interaction techniques.

2.3.1 Asymmetric interactions techniques

In this section techniques from Malik et al. [8] are introduced where the left hand supports the right one. As seen in the daily life users tend to support their right hand by their left hand (if you are right handed). For example, you hold the glass with the left hand and do the pouring by your right hand. Also then using a desktop pc you use your left hand on the keyboard to support the the right hand. For example you switch fast between open windows with the Ctrl+Tab short-cut. These

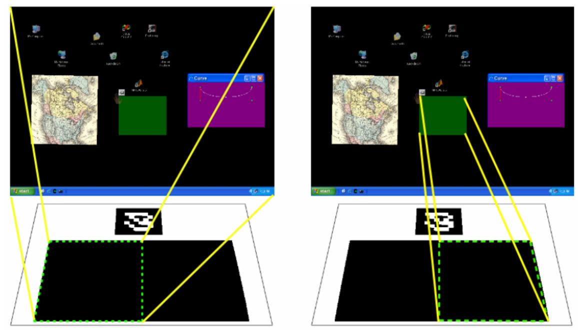


Fig. 9. Mapping of the area for the left hand (left) and touchpad mapping for the right hand (right) [8].

techniques are introduced by Malik et al.:

- Coarse positioning

  As seen in figure 9a the touch input surface is divided into two parts. The left part is mapped to the four corners of the entire display so the left hand can access all the elements on the display. But regarding to the big screen fine positioning will be difficult.

- Workspaces and fine positioning

  The next concept enhances the fine positioning. For that the left hand is tracked so that a green-coloured box can be attached to the position of the index finger defining a workspace. This workspace determines the detail space, which is mapped to the right part of the input surface. Now the right hand can perform much finer actions due to the higher resolution. The bimanual interaction offers quick movement of the cursor to any position and in the same glance accurate interaction with the elements on the monitor.

  To my mind it is a very good consideration to utilise the natural handedness of people like shown in these two techniques. Especially for very big screens, keyboard and mouse can not offer this fast switching between resolution.

- Selecting and moving of single objects

  As soon as the index finger of the right hand touches an object on the surface it becomes selected and can be dragged around by sliding the finger over the surface. Furthermore the hole workspace with the selected object can be moved around by the left hand and "thereby allowing the object to be coarsely placed anywhere on the screen quickly, but without interfering with any precision movement being carried out by the right hand [8]".

- Selecting multiple objects

  In classical interfaces the user is required to stretch a box around the elements with the mouse or select them individually. For Malik's et al. new approach it is necessary that the hand tracking is performed by a video camera from the top. With this ability you select an element by grabbing it with your hand. After this the closest object to the centre of the right hand disappears and is queued. Doing this again and again makes it possible to select multiple objects. To paste a queued object to the interface the user splays out all fingers of the right hand. When doing this, a picture of the hand is shown on the display with the icons of the previous selected elements appearing on each finger. Now a simple tap of a finger will paste the element to the tapped position.

While being very innovative this method has as a big drawback with the necessary of a camera based finger tracking. So only fixed working places can offer such capabilities. For future research I suggest a gesture for cutting out objects by placing all finger tips around the object and then slide down and out of the touch surface. The correspondence to the real world task is wiping or taking an object from the desk.

- Resizing the workspace

  By default the workspace has the granularity of a pixel so the corners of the tracking part for the right hand are mapped to the corners of the green box. But it might be desirable to change the size to adjust the mapping to the current necessary accuracy. For this purpose you just have to stretch the index finger and the thumb apart for up-sizing or to approach the index finger to the thumb for down-sizing.

- Pinned workspaces

  If there is a very large screen like a power-wall, the working regions may concentrate on just two spots on the display. If this spots are distant to each other the use has to move a long distance. In this case it may be interesting to pin down a workspace in size and position and open a new one. To accomplish this the user double taps with the left hand to a position. By default the right hand is now trapped in the pinned workspace. By pointing with the left index finger a new workspace can be created and dragged as before. If this new workspace overlaps the old workspace, the mapping for the right hand surface area changes to the new one. So several workspace with different sizes are possible.

2.3.2 Transitioning from asymmetric to symmetric interaciton

While asymmetric interaction will fit for many tasks, the user might wish to work with both hands within a single workspace in some cases. To do so, he slides with all his fingers of the left hand to the bottom-right corner of the surface. Now symmetric interaction tasks can be performed. To change back to the asymmetric mode the left hand slides with all fingers to the top-left corner. Unfortunately Malik does not suggest interaction techniques to boost the user performance in symmetric interaction. Regarding to this I want to suggest some interaction techniques from Benko et al. [2] for future research. I think that these techniques offer context aware options like the right mouse button.

Benko et al. [2] introduce several selection techniques for multi-touch screens. Through they assume them for direct touch input it would be worth to have a look at them and mark them for future research in indirect touch interaction because the relatively high error rates, arm fatigue and the lack of precision still prevent the widespread adoption of touch input for general computing devices. Especially then using both hands in a symmetric way and in combination of the tracking state following techniques will give a richer interaction space.

- Dual finger stretch

  The dual finger stretch will allow to stretch a small part around the right index finger. To do so the left index finger touches near the right finger and then drags away from it as seen in fingure 10. Concurrently the right finger is now able to move within this zoomed area with much higher precision which is the main advantage of old zooming techniques. Lifting of the left finger removes the zoomed area. The only limit of the zooming is the size of the touch surface. Usually a change in aspect ratio can be achieved by up to 10. Striking example for a working task there this come to handy is picture manipulation. Using this technique pixel sized action are possible.

  I also would suggest another variation that the zooming is just done by the index finger and the thumb of the left hand. Moreover its possible to move the zoomed area by moving the left hand which is quite similar to the asymmetric technique of coarse positioning.

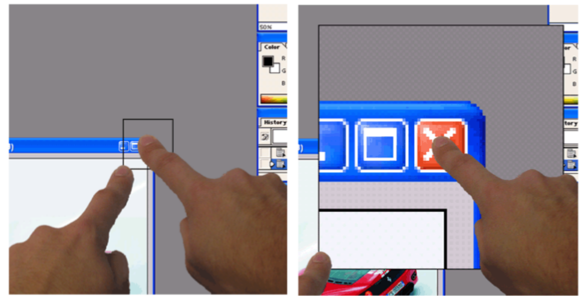


Fig. 10. Dual Finger Stretch technique for zooming a selected part of the screen: Initial selection of the zooming area (left) and zoomed situation where the primary finger can perform actions easier (right) [2].

- Dual finger x-menu

  The next technique can offer a similar menu than the right mouse button. Whenever the left index finger touches the surface a circular menu is invoken around it. Moving the finger across an item selects it. In this case (*as seen in figure 11*) the menu offers options for zooming, snapping and several settings for different relative speeds. "Normal mode moves the cursor with the same speed as the primary finger; the two slow modes reduce the speed of the cursor by a factor of four and ten respectively, while freeze mode "freezes" the cursor in place, disabling any cursor movement" [2]. The snapping mode removes the offset of the cursor immediately and moves it to the current finger position of the right hand. The magnifier brings up a zoomed (by factor 2) picture of the area under the right finger in the center of the x-menu. This is quite useful then the finger would occlude the display but due to we assume indirect touch interaction this options can be removed.

  I have a few reservations about the way the x-menu is invoken. The single touch of the left hand could have some better action attached so why not touch the surface with two very closed fingers? So the left index finger could move around the workspace and offer in need the x-menu. Moreover, it would be an improvement if the menu were context away. In this way it would be a very good technique to replace clicking with the right mouse button.

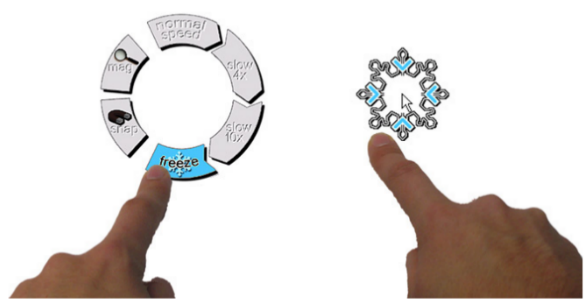


Fig. 11. Dual Finger x-Menu which offer different options. Currently the freeze option is selected [2].

- Dual finger slider

  "Encouraged by the possibilites of the different interactions modes of Dual Finger X-Menu [...] [Benko et al.] developed the Dual Finger Slider technique, which incorporates the menu's most useful features, but simplifies and steamlines the overall interaction" [2]. The user can perform touching tasks with his right hand as usually, but can always adjust the pointer speed by moving the left index finger to or away from the right one. In detail the graphical circular menu and the option magnifier are removed. Now the natural ability to gauge the distance from the left index finger to the right index finger shows the speed ratio of the mouse pointer. Here also a peripheral field of view is necessary. Moving the finger to far away triggers the snap option and the cursor jumps back to the right index finger.

This were some interaction techniques for direct multi-touch input. The goal is to combine them in such an manner that general computing devices with touch input become true.

3 Conclusion

We have seen that direct touch has drawbacks especially for classical work environments. Indirect touch interaction can be a solution for these drawbacks but it is still under research. Gilliot et al. did some first research about the device size and the peripheral field of view and have found important facts. I think that it is very important for the user to have the input surface in their field of view so they can match the objects on the monitor easier to the right spot on the touch surface. Moreover if the aspect ratio of the input surface is not the same as the display ratio user perform worse because they can not map the elements to the surface. Also it's good to keep in mind, that the display size doesn't affect the user performance so interaction with big wall mounted screens like the powerwall is generally possible.

Secondly the research of Voelker et al. has shown that the re-introduction of the three state model (see figure 6) with the tracking state is essentially for a good user performance. While the gesture based techniques may be applied to most touch surface technologies I see the pressure based techniques being superior to them. Whether to use the pressure quasimode or the pressure switch method should be subject of further research. Also Benko's SimPress Clicking should be examined in this particular case. When doing research Voelker's postulated design considerations are a good criteria to evaluate the outcomes. Especially being aware of the current state and not slipping unintentionally between the states is crucial.

Obviously bimanual interaction with all fingers can offer more input possibilities. The techniques investigated by Malik et al. give a good starting point for future research. Especially if the tasks demand a high accuracy the workspace concept may offer the needed precision. But I have some reservations about being able to pin down them and open a new workspace because in Malik's concept you can't jump with your right hand between the workspaces easily. Mapping all of the open workspaces to the surface side of the right hand can be an opinion, but this will interfere with Gilliot's findings of the aspect ratio.

To my mind symmetric touch interaction may offer additional accuracy but is not covered by Malik's research. So this is up for future research. A good start can be Benko's work in which he introduce several symmetric interaction techniques for direct touch input.

Owing to all these considerations it becomes plausible that none of the suggested methods can be alone the solution of the future working environment. All of them have their advantages, but also their drawbacks. At the end it will be the combination of several techniques to evolve the general desktop environment and offer a good user performance.

References

[1] P.-A. Albinsson and S. Zhai. High precision touch screen interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 105–112, New York, NY, USA, 2003. ACM.

[2] H. Benko, A. D. Wilson, and P. Baudisch. Precise selection techniques for multi-touch screens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 1263–1272, New York, NY, USA, 2006. ACM.

[3] X. Bi, T. Grossman, J. Matejka, and G. Fitzmaurice. Magic desk: Bringing multi-touch surfaces into desktop work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2511–2520, New York, NY, USA, 2011. ACM.

[4] W. Buxton. A three-state model of graphical input. In *Proceedings of the IFIP TC13 Third Interational Conference on Human-Computer Interaction*, INTERACT '90, pages 449–456, Amsterdam, The Netherlands, The Netherlands, 1990. North-Holland Publishing Co.

[5] C. Forlines, C. Shen, and B. Buxton. Glimpse: A novel input model for multi-level devices. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pages 1375–1378, New York, NY, USA, 2005. ACM.

[6] C. Forlines, D. Wigdor, C. Shen, and R. Balakrishnan. Direct-touch vs. mouse input for tabletop displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 647–656, New York, NY, USA, 2007. ACM.

[7] J. Gilliot, G. Casiez, and N. Roussel. Impact of form factors and input conditions on absolute indirect-touch pointing tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 723–732, New York, NY, USA, 2014. ACM.

[8] S. Malik, A. Ranjan, and R. Balakrishnan. Interacting with large displays from a distance with vision-tracked multi-finger gestural input. In *ACM SIGGRAPH 2006 Sketches*, SIGGRAPH '06, New York, NY, USA, 2006. ACM.

[9] W. S. Marras. Basic biomechanics and workstation design. *Handbook of Human Factors and Ergonomics*, 3:340–370, 2006.

[10] M. R. Morris, A. B. Brush, and B. R. Meyers. A field study of knowledge workers use of interactive horizontal displays. In *Horizontal Interactive Human Computer Systems, 2008. TABLETOP 2008. 3rd IEEE International Workshop on*, pages 105–112. IEEE, 2008.

[11] D. T. Pawluk and R. D. Howe. Dynamic contact of the human fingerpad against a flat surface. *Journal of biomechanical engineering*, 121(6):605–611, 1999.

[12] D. Schmidt, F. Block, and H. Gellersen. A comparison of direct and indirect multi-touch input for large surfaces. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I*, INTERACT '09, pages 582–594, Berlin, Heidelberg, 2009. Springer-Verlag.

[13] S. Voelker, C. Wacharamanotham, and J. Borchers. An evaluation of state switching methods for indirect touch systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 745–754, New York, NY, USA, 2013. ACM.

Interactive Non-Expert Information Visualizations and their Evaluation Beyond Time and Error

Johanna Fulda

Abstract— In this paper we look at the current state of interactive information visualizations and how they can be evaluated beyond time and error. We focus on non-expert visualizations that contain interactivity for the user and can be found in the environments of news websites, museums and public spaces. For each environment particularities and examples are illustrated and the different categories of interaction described, including for what purposes they can or should be used. For evaluating those graphics we look at the advantages and different use cases for "Lab Studies" and "Into-the-Wild Studies", and explain why qualitative and quantitative methods both are equally valuable. Furthermore we point out three aspects that should be considered when evaluating those visualizations: 1) The inclusion of the building process into evaluation, with user testings as well as expert analyses, 2) Rating the discoveries a user makes during exploration, and 3) The advantages of long-term studies.

Index Terms—Infovis, Evaluation, Visualization, Interactive Information Visualization, Evaluation methods, Non-Expert Information Visualization, Lab Studies, "Into-the-Wild" Studies

1 Introduction

If we consider previous research in the evaluation of information visualization (Infovis), there are two main components being considered, which are representation and interaction [39]. There is still more focus on evaluating the representational part [11], but with the increased dissemination and the facilitated possibilities to create interactive visualizations it is becoming more important to also focus on the interaction part and try to understand how people use it and how they gain insight. Of course these two parts go hand in hand, because also the interactive visualization needs to have a clear and understandable representation to make it usable. The interaction part however is still underrepresented in scientific research papers so far [39].
To first understand the advantage of the opportunity to interact with infographics, Endert et al. explained how it helps understanding complex and extensive datasets [12]. That is because the user can visually explore data, make decisions independently and navigate through the available dataset corresponding to his interests. Especially for the interactive "non-expert" graphics considered here, also the playful aspect is important, because the user has to be engaged and motivated to deal with the graphic. So the question arises how we can evaluate the quality of those interactive visualizations. For example they can either contain a lot of useful information but being ignored by the users because they don't catch their attention, or they are just not recognized as something to interact with and offering information - that can particularly happen with installations in public spaces [16]. Another possibility is, that it invites the users to play, but misses the informational part because the information isn't easily accessible or it just doesn't provide a lot of it. To regard all these circumstances for the evaluation, it isn't enough to only look at the time and error rate in a user study, but we also have to consider how likely it is that the graphic catches the users' attention and prevails upon them to interact with it, how much insight a user gains and how he understands using the provided possibilities of interaction. Therefore we look at established evaluation methods and their benefits. We also point out some aspects that should additionally be considered when evaluating interactive non-expert visualizations. Before that we define the addressed visualizations more precisely, present some examples and show different categories of interaction.

• *Johanna Fulda is studying Media Informatics at the University of Munich, Germany, E-mail: johanna.fulda@campus.lmu.de*

• *This research paper was written for the Media Informatics Advanced Seminar 'Special Aspects of Usability', 2014*

2 Considered visualizations

Information visualizations can be used in a lot of different areas. For this paper we narrowed down the target group and decided to focus on interactive "non-expert" visualizations. Those who can be found in three different environments. The following explains the details and shows some examples.

2.1 For "non-experts"

Many information visualization systems are considered to visualize a very specific research field and thus help experts doing analyses of their highly specialised work. But there are also use cases for Infovis systems that aim for explaining a general topic, where the target group is people with various backgrounds and different fields of expertise and interest [28]. Most of them are thus no experts in the specific field of the visualization. These "non-expert" visualizations are therefore rather for giving an overview of something and do not intend to dive into the topic too deeply. Even though the interactivity enables further exploration, it has to be assumed that the user is new to the topic and explain everything from scratch.
In general the use of those visualizations is voluntary and they often are just supplemental offerings. That means it first has to get the users attention and look interesting enough to motivate further investigation. There are a lot of factors influencing that decision, for example time constraints, familiarity, mood, age [18]. If the visualization attracted the user, the possibilities of interaction should be easy to discover. They should either be recognizable from previous experience and correspond to a "usual" way of interaction - as known from web applications or native apps - or provide the user with a challenging or creative task [18]. But all interactive elements have to be identifiable, more advanced interaction methods should be explained shortly. If it looks too complicated it might scare off the users, so there has to be a trade-off between looking interesting and being manageable [6].

2.2 Environment

These non-experts visualizations occur in many different environments. To narrow down the field of application, we want to look at three different cases, as there are news websites, museums/exhibitions and public spaces. Below, all three environments are explained in more detail:

2.2.1 Interactive graphics at online news websites

Since Internet technologies, such as browsers or native apps offer many possibilities for interaction, it challenges the traditional "one-way directional flow of news" [9]. The readers can choose themselves what they are interested in und don't have to follow the guidelines

from the news producers [8, 17, 35]. They can choose from endless offers and even get the chance to participate interactively in different ways. One kind of interactivity is interactive graphics, which can be associated with an article, provide additional information or give an overview over a certain topic. There are several flagship examples from big online newspapers, like *The Guardian* or the *New York Times*. Together with the new species of data journalists they try to realize big interactive data projects more and more often.
In contrast to the traditional newspaper, that people often read calmly for example during their breakfast or while riding a train, the purpose of reading online news is mainly to get informed or updated very quickly, for example during a short break at work - but it can happen several times a day [17]. When online news users enter the landing page of the website, they find an overview with a huge selection of articles. The goal of the users is to find something that is interesting for them. Again in contrast to the traditional newspaper, where the reader at least scans the content and layout of most articles by flicking through the pages, the online news only show a small teaser of the article or even just a headline to invite to further reading. That means that in the end most texts there will never be opened [17]. So the first challenge for the interactive graphics is to be realized at all. They have to catch the readers attention and motivate them to follow the link. The second one then is the understanding of the graphic. The readers have to understand what the graphic shows, how they can interact with it and what advantages they get by doing so. And all this should be visible without the need for an extensive explanation [8].

Figure 1 shows an example of an interactive graphic at *sueddeutsche.de* (the online presence of the *Süddeutsche Zeitung*), which was developed together with *OpenDataCity* - the ”Zugmonitor”. It visualizes data on a map from the German Railway Company in real time and gives an overview of all trains and indicates those who have a delay. The view can also be predated, and show data from a former day. The data is obtained through data scraping from the train company’s website, thus it unfortunately is dependent on their data structure and won’t work anymore when the company decides to change it (as it happened in autumn 2013). It still is a good example for the possibilities of interaction a news website visitor can have.

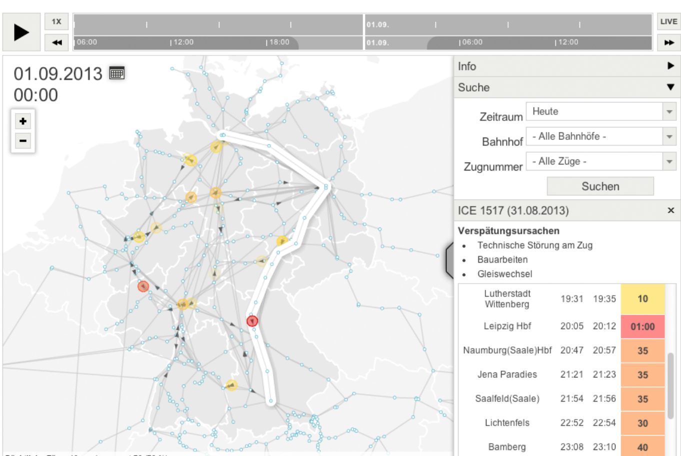


Fig. 1. Screenshot of the ”Zugmonitor” from sueddeutsche.de [33].

2.2.2 Museums and exhibitions

Increasingly augmented with digital technology [18] museums often offer interactive Infovis systems to their visitors. Those provide deeper insight or background information to the exhibited pieces. Care has to be taken that people visit museums for very different reasons, as for example to ”add to their specialized knowledge”, or for ”an entertaining and educational experience” [15]. That means on the one hand we cannot assume that visitors have prior knowledge in a certain topic and thus the installation should provide introductory information. On the other hand it should also be interesting for those who already have specialized knowledge and offer more profound information. Furthermore it is not very likely that a visitor will use the installation more than once [18] so it has to evoke their curiosity immediately as well as be understandable intuitively. But at least we can ascribe the visitors a certain interest for the exhibition.

As examples the *Austrian Technical Museum* in Vienna had an installation at their exhibition medien.welten (in summer 2013), where visitors could do calculations like on a real Abacus. The system evaluated the input and offered feedback and instructions if necessary [18]. Figure 2 shows another example, the ”The Virtual Fraunhofer Spektralapparat” at the *Deutsches Museum* in Munich. An interactive 3D model is displayed next to the original piece. The visitor can turn it around in every angle and gets additional information about the functionality of the Spektralapparat, like which part does what and how. This approach brings the visitor way more insight than if he was just looking at the model [21].

Fig. 2. The Virtual Fraunhofer Spektralapparat at Deutsches Museum Munich [21].

2.2.3 Public spaces

Public spaces are cluttered with displays nowadays. The vast majority contains passive content like advertisement, directions or restraints. Digital screens sometimes also display location-dependent information, such as the time one has to wait for the next bus or train, the local news in an underground station or redirections because of traffic jams on the road [10]. There are a few examples of visualizations that include the urban environment and shows information that is generated by the people around [23]. In Copenhagen and Aarhus for example, there are some ”Cykelbarometers” placed next to much-used cycle ways. They count and display the number of bikes passing by through the day. In New York the ”National Debt Clock” shows the American government’s debt on a huge display and also indicates how much that is for each American family [23]. (In 2008 it ran out of digits [4].)
Only very few attempts have been made for installations where people have the possibility of ”hands-on” interaction in public spaces, partly because the displays still are a case of cost and maintenance, but above all because it is expensive and time-consuming to offer interesting and interactive content that matches all the requirements these installations have [2].

2009 the MIT SENSEable City Lab developed a design for a futuristic bus stop as seen in figure 3. The ”EyeStop” that is partially covered with touch screens and sensors, offering people to plan their trip (through entering their destination and getting a map with the shortest way), to see where the bus they are waiting for is right now or to write or read local announcements (and much more). Although it is not implemented outside in a real public space yet, it surely is a showcase project for ”the potential of next-generation urban transportation design” [25].

Fig. 3. EyeStop, developed by MIT’s SENSEable City Lab [25].

One less futuristic example is the installation ”Vote With Your Feet” from Steinberger et al.. A ”hyperlocal public polling tool for urban screens” [13] that is meant to engage people in thinking about certain topics and to show the community’s attitude towards it. With two buttons on the ground, the participant can answer questions with ”Yes” or ”No” and thus be part of a survey.

But the few interactive installations in public spaces seem to have additional problems with getting noticed and accepted. Especially in cluttered public spaces it is at first challenging to get people’s attention and to make them pause for something they don’t know and don’t see the immediate benefit of [24]. Many objects in public spaces may not be noticed at all or at least not with the main focus of a passing person. Sometimes they are seen unconsciously or as secondary task, while doing something else, which means people are paying only limited attention to these objects [32]. Furthermore the common human-computer interaction (HCI) findings for how people interact with interactive displays seem to differ a little for public spaces, as there is the factor of being in public and seen by others [24, 13, 34] and the factor that the system has to be realized before the interaction can take place [24] - HCI mostly assumes that the user already knows that there is something to interact with.

3 Categories of Interaction

In his book ”Visual Explanations: Images and Quantities, Evidence and Narrative” Tufte lists 58 different methods of interaction to ”winnow the wheat from the chaff and separate the sheep from the goats”, those are for example ”pair, merge, harmonize, synthesize, focus, organize” [37]. To narrow those down a little, Yi et al. made one persuasive categorization with seven methods. Those are: Select, Explore, Reconfigure, Encode, Abstract/Elaborate, Filter, Connect [39].
For the visualizations we observe here, not all of those seven methods seem equally important, because for the non-expert user (as defined above) the interaction should primarily be simple to lower the barrier to get started. If there is too much preparation necessary, it is very likely that people abandon the interaction.

As Table 1 shows exploring seems to be the most prevalent method of interaction for the non-expert interactivity. It enables the user to change to a different subset of data. Sometimes exploring reminds more of flipping through a book than being in a modern digital environment, where more fancy methods could be applied. But to also encourage inexperienced users to interact with digital devices it’s reasonable to use that method [19].

To show the applications for all the other interaction techniques, the seven methods are defined shortly and explained how they are used in the examples from Table 1:

3.1 Select

Enables the user to mark one item as interesting and thus lets him keep track of a transformation or get additional information for that item. In the example of the ”Zugmonitor” the user can click on one train to highlight the route it is on. It simplifies tracking one specific train without having to stare at it to not get lost in the clutter. Furthermore it lists all the stations that particular train is stopping, the scheduled time and the potential delay.

3.2 Explore

As already mentioned Explore lets the user change the displayed subset of data. It doesn’t necessarily change the whole view but can also just let some additional data appear and some previous data vanish. In the mentioned examples in Table 1 the interaction mostly changes the whole view.

3.3 Reconfigure

The spatial arrangement can be changed through reconfiguring the view, for example to be able to see more details and less clutter at a specific area of interest. Hidden information can thus be revealed and connections or rankings seen more clearly. For listings reconfiguration is used very often, to sort lists after specific criteria. In the example of the New York Times graphic, where the World Cup players are shown in different sizes depending on the number of mentions in Facebook posts during a day, the user can also reconfigure the view and either order them by the country they are playing for, by their name or by the visualized number of mentions. In the examples for the environment of a museum the reconfiguration is often used for 3D models, to enable the user to virtually turn around an object and observe it from any angle.

3.4 Encode

Encoding can be used to change the fundamental visual representation, like colours, shapes, the kind of representation - for example change a bar chart to a scatterplot. This technique doesn’t appear too often in the considered Infovis systems, therefore there is also no example for it in Table 1. The reason may be that encoding requires the user to already have a deeper understanding of the represented dataset, to be able to purposefully change the representation to a more suitable one for that specific dataset. It could surely also be used for playing around and thus accidentally discovering new features, but it probably is more useful for expert visualizations.

3.5 Abstract/Elaborate

Especially for big datasets it is impossible to show all information in one screen and very often it is also not necessary, because the user only wants an overview without going deeper into detail or he is only interested in a single aspect. Abstracting and Elaborating the infographic allows the user to adjust the degree of abstraction - to either go deeper into one aspect or go further away to see more of the overall image. At *The Guardian*’s ”Woman’s Rights” graphic, the reader first gets an overview of all the continents and their averaged data. If he then selects one continent, the graphic goes deeper into the data of that specific continent and shows details for the single countries inside. Here the colours and their saturation indicate the positioning of one country in the seven represented categories. If the user again selects one of those sections, a listing for the legislation behind that category is shown next to the graphic, displaying a tick or a cross for being in the country’s law or not. In that graphic there are three layers of abstraction. Many more can for example be found on geographical maps where data can either be seen for the whole world or alternatively for one street inside a city (and many layers in between).

Environment	Name / Place	Underlying data	Tasks	Explanation
Websites	**Zugmonitor** Süddeutsche Zeitung (zugmonitor.sz.de)	data scraping from Deutsche Bahn website	Select, Explore, Filter, Connect	Visualizes long-distance trains in Germany and their delays. Live and historic. (Currently out of order since the German train company changed their data structure in autumn 2013)
	Top World Cup Players on Facebook, Day by Day New York Times (nytimes.com/interactive)	Number of mentions of players through Facebook API	Explore, Reconfigure, Encode	Pictures of the most discussed World Cup players are sized corresponding to how often they were mentioned on facebook posts (for each day of the World Cup 2010)
	Women's rights, Country by country The Guardian (theguardian.com)	Excel Sheet with data collected by World Bank and UN	Explore, Reconfigure, Abstract/Elaborate	Illustrates women's rights across the world. Details for single countries in seven categories and the legislation behind them (February 2014)
Museums/ Exhibitions	**Fraunhofer Spektralapparat** Deutsches Museum	3D model of Spektralapparat and associated explanations	Explore, Reconfigure	Digital 3D model next to the original artefact to let visitors discover the piece more detailed
	Abacus Technical Museum Vienna	Exercises generated by museum	Explore	Digitally-augmented Abacus that lets the visitors do calculations on it and offers feedback and instructions
	360grad Electric BMW World in Munich	3D model plus additional information	Explore, Reconfigure, Abstract/Elaborate	Shows a BMWi model and background information with slick animations
Public Spaces	**EyeStop** MIT SENSEable City Lab	APIs to public transport, news, etc., plus data from own sensors	*Everything possible*	Solar powered future bus stop that should also be used as a community gathering space
	Cykelbarometer Copenhagen, Denmark	Generated by people	Other (more participation than interaction)	Displays the number of how many cyclists already past that point that day
	Vote With Your Feet Brisbane, Australia	Generated by people	Select	Public polling tool installed at bus stop asking people Yes/No questions. Answers are given through two buttons on the ground

Table 1. Examples for different environments

3.6 Filter

Filtering means the user can add or remove criteria for the displayed data and thus only see the matching information. He can for example select a range of time, because it is assumed that in that range something interesting happened. All the other information is not interesting at that moment and thus doesn't have to be displayed. In the "Zugmonitor" the user can filter the data for specific dates or time frames or he can choose one specific train station or one specific train number and get information about delays only for trains that match that criteria.

3.7 Connect

Connecting representation can either be used when one dataset is visualized in different ways in more than one view - to highlight one selected item also in all the other views. Or it can be used for highlighting similar items to show connections that otherwise would have remained unseen. Again in the "Zugmonitor", if the user clicks on one train its route is highlighted and thus shows the connected stations. That gives the user additional information because with the many routes overlapping and cutting across each other, it is impossible to recognize one specific route if it is not highlighted.

As table 1 shows not every kind of interaction fits into those definitions. Especially in the environment of Public Spaces it's not necessarily only touchscreens the user can interact with, but different objects that for example augment the visualization, make it more special or artistic.

4 Common Evaluation Methods

There are many well-established and valuable methods for the evaluation of information visualizations. They are mostly well documented and the creators of the study can use guidelines and checklists to prepare their tests. The types of studies reach from simple task solving exercises to more complex observations of unknowing users. There are quantitative and qualitative methods that can either take place in a lab or outside in the settings where the applications will actually be used when finished. Advantages and disadvantages of both environments will be explained and why assertors of qualitative and those of quantitative methods don't have to fight any longer. But first we discuss why we need more elaborate approaches than "the time and error approach" at all.

4.1 Problems with the time and error approach

If there is a classical way of evaluating information visualizations, then it probably is the one where a test user is asked to deal with a certain task and then the observer watches how long it takes him and how many errors are made in the meantime [30]. This approach is sufficient for simple visualizations, but with more and complex data and many different possibilities of interaction it becomes less suitable. Furthermore tools that visualize big data are mainly made for "innovation and discovery" [31] so it doesn't make sense to let the user do one specific exercise, but rather observe how he is trying to gain insight and discover the data. To "quantify the quality of a visualization system" [1] we thus need methods beyond time and error.

4.2 Lab Studies

It is very common to evaluate new things in the environment of a lab, because there the researcher has the necessary equipment and all the test users are in the same situation when they do the tests. Also ethically it is flawless because the users know that they are part of a study and can thus not feel spied on afterwards.

4.2.1 Usability testing

The classical usability test requires three to ten test users that are given a set of typical tasks. The observers watch them to see where they run into trouble and write a report afterwards with all the identified problems, possibly ranked by importance. Then the product is being improved and a new usability test created. Through these many iterations it is possible to improve the understanding and the operationality of a product. And it can respond to the needs and wishes of the end user in a very effective way. Disadvantages are that the environment is different than the one the system would usually be used in. Also the devices may be different from the ones used normally. Improvements can be to relocate the tests into the work environment (including all the distractions and interruptions one normally has there) and to allow the test users to utilize their own devices [31].

4.2.2 Questionnaires, Talking Aloud and Thinking Aloud

Additional to only looking for obstacles the user is facing, one can ask the test person a lot of questions afterwards, why he decided to do it this or that way. More live and unfiltered feedback one can get by asking the user to say everything he is doing out loud, that also reveals the consciousness about decisions and if everything is understood as it was supposed to. The "Thinking Aloud" method requests the user to talk even more. He has to speak out everything he is thinking while finishing the exercise. That can reveal indeterminations of what to do next or show how certain elements are received and if they are recognized as what they are supposed to be. But because the talking task demands a lot of concentration and users also tend to pay more attention to navigation problems than they normally would, those tests can not at the same time be used for testing the time needed for completion and thus the efficiency [38]. They are also highly subjective and can't be evaluated with an algorithm or some mathematical formula but have to be assessed individually (see 4.4).

4.2.3 Cognitive methods

To not only rely on what the test users verbalize, there are methods where the accurate behaviour can be tracked and thus even unconscious activities can be noticed. Methods are for example "Eye Tracking", measuring the brain activity or even more physiological responses like muscle activity. All these measurements need quite expensive hardware, so it can't be expected to have them available for every project. If they are available, again the final evaluation is another big challenge, because all the recorded signals have to be interpreted. Therefore the researchers need special knowledge in neuroscience or have to collaborate with experts. It also makes sense to already involve experts during the design phase of the study, the results may be less cryptical then for non-neuroscientists - but still need some extra knowledge [3].

4.3 "Into-the-Wild" Studies

For the interactive non-expert visualizations we defined above, the previous described methods are not sufficient, because they only consider the last of the three phases. The first important phases of the evaluation of those visualizations is, that they have to be discovered and approached, before the interaction finally takes place. So we can not just invite test persons into a lab and place them in front of the system, but we want to know if untaught people do realize the existence of the systems at all and if so, if it makes them curious enough to actually approach them. That means we actually have to do tests outside "in the wild" to explore people's behaviour.

4.3.1 Awareness and Motivation

There are at least three phases that have to be gone through for a successful interaction [24, 16] - some literature even mentions more phases (for example Michelis et al. name six in their "Audience Funnel" [22]), but those are the ones that seem to be most important for us:

- **Attention**: The first step is to be realised by passing people
- **Motivation**: The phase of understanding what it offers and deciding if it is worth trying
- **Interaction**: The actual phase of interaction and gaining insight

As mentioned above, we have to leave the lab environment to be able to evaluate the first two phases. For news websites it may even work to explore the user's behaviour in a lab, if the user doesn't know about what is there actually to discover, but the testing environment can bias the behaviour in that way, that the user is more attentive than he would normally be, when browsing a news site. So we want to find out, if the visualization is being discovered without the user having any expectations. That means we let unknowing people approach and interact with the installation and can start asking questions afterwards.

4.3.2 Observation and Interviews

In the environment of a museum Hinrichs et al. chose a "qualitative ethnographically oriented study method" as described by Blomberg et al. in the book "Participatory Design - Principles and Practices". People were informed through a sign that a study was being conducted in the room to harm nobody's privacy. Even though they didn't use video or audio recordings and only took notes of their observations. The observer was also sitting quite far away from the installation, so they hoped there was no interference with the people's behaviour. After interacting with the object, the visitors were asked to fill out a questionnaire concerning their experience with the installation and if and how they gained insight through the visualization. Afterwards the researchers had to analyze all the notes and discoveries. Therefore they used the "open coding method" (the term was characterized in 1989 in the context of Social Sciences [5]) which means creating a category system based on the results of the observation. The categories, as many as necessary, are generated freely and afterwards grouped together and given more abstract labels, then grouped together again and so on [7]. The results then showed different types of visitors, different behaviour if they were alone or in a group and how long and intense they interacted with it. They concluded that the motivation for approaching the installation was "the display technology, the visual appearance of the visualizations, and seeing other people interact with it" [15].

4.3.3 Hidden cameras

It often makes sense to capture the situation with a camera, to enable looking at it again afterwards. There are some rules the researchers have to stick to, to not infringe the test person's rights but if they keep those conditions it is a useful tool. It makes it even possible that the observers are at another place during the test phase, to not influence the behaviour of the passing people. For example Steinberger et al. used that method to evaluate their interactive bus stop. Like in the previous example they also observed the people (the ones using the object as well as the people watching) and tried to interview them after they finished their interaction with the object. Here the motivation for approaching the installation was either an interest in the displayed topic or it was seen as a nice occupation while waiting for the bus. Due to the fact that it offers a very simple task (one Yes, one No button) it was perceived as very easy and understandable [13].

4.4 Qualitative vs. Quantitative

All these observation, interviewing people or letting them fill out questionnaires entail a long post processing phase. They involve a big interpretive part and collection of empirical data, sometimes even philosophical questions. These qualitative methods ”seek to make sense of personal stories and the ways in which they interact” [27, pp. 1]. In contrast to that there are the quantitative methods, which use numbers and statistical methods to produce results that are generalizable. ”In quantitative research, the researcher’s role is to observe and measure, and care is taken to keep the researchers from ’contaminating’ the data through personal involvement with the research subject” [27, p. 6]. In many older literature about research methods, scientists were almost fighting about which of those methods is better. Positivists ”regard the world as made up of observable, measurable facts”, whereas the interpretivist ”portrays a world in which reality is socially constructed, complex and ever changing” [27, p. 8-9]. Nowadays scientists agreed largely on both methods being useful for specific needs. And it is also common to use them both, for example enriching collected numbers with user interviews or to also make a survey with predefined answers during a qualitative study [36, p. 3-6].

For the considered Infovis systems the qualitative approach often is more convenient. The creators want to know how people perceive the objects and what exactly motivates them to actually approach them or rather not. Also because it is a not too established field of Infovis, they want to get new insights, suggestions and ideas, to keep improving the systems. And those insights may not be caught by quantitative methods. An exception can be the evaluation of online graphics. Because here the persons responsible often don’t want to spend too much time for evaluation, so they draw back to the data that is generated for their website anyway. That is number of visitors, number of clicks, time they remained on a page, and if available some demographic values (age, gender, field of interest), where they come from (geographically and from which previous website), if they are new or returning visitors and so on. With that data they can see if a visualization is very popular or if it seems to remain rather undiscovered. The data about how long someone spends with it reveals insight too, but the reasons, the exact behaviour, and understanding cannot be caught through that data. A tool for getting qualitative feedback on websites (without having to prepare a study) is the comment function they often offer. People who are particularly excited about the article, or feel the need to contribute something get the chance there - however if that feedback is used to improve future graphics cannot be answered explicitly.

Many of the portrayed methods are established and well documented and can deliver qualitative as well as quantitative data. But especially for the considered interactive non-expert visualizations there are some ideas for additions to common evaluation methods that still could improve the results. In the following chapter we take a closer look at those ideas.

5 Required features

The previously portrayed methods are used very often and there are also many detailed instructions available. ”A method of analysing interview transcripts in qualitative research” [7] for example guides through all the steps of a qualitative evaluation, the book ”Blending qualitative and quantitative research methods in theses and dissertations” breaks it all down and offers a whole ”Catalogue of Methods” [36]. Still there are some features missing in current evaluations, three suggestions are described in the following. The inclusion of the building process and the measuring of discoveries can be performed in lab study environments, Long-term studies would then take place in the final environment of the system.

5.1 Include building process

There are two aspects that should be evaluated already and repeatedly during the creation process of the interactive visualization. That is the interaction (usability and understandability) and the appeal (that includes aesthetics and the user’s experience). Especially usability and experience overlap very often and the effects of aesthetics are sometimes only perceived unconsciously by the user. That is why next to testing the prototypes with users it also makes sense to include interaction design experts in the building process [14]. Through their experience and knowledge of guidelines they can help to avoid making mistakes that have been made before. Both suggestions are very time consuming and expensive but can avoid a lot of post-treatment.

5.1.1 User testing

As mentioned earlier the non-expert visualizations have to be very easy to understand. Some projects have been portrayed that have the simplest possible interaction methods (for example only 2 buttons). But the bigger and more complex the data gets, the more possibilities of interaction could be used. Therefore a balance has to be found between using as many handy methods as possible and still being usable intuitively. But very often developers and designers lose their neutral view during a project and tend to overlook potential problems, because it seems just very clear to them. As idea for evaluating the balance between interesting and simple, it would make sense to already include the building process into the evaluation [20]. As described for the usability-tests, one could regularly invite few test users to see if they understand the current possibilities of interaction and based on that continue the development. That would avoid having something totally magnificent in the end that people won’t use because they don’t understand it. As described by Hinrichs et al. ”a visually appealing information visualization can be experienced negatively if it is hard to explore due to awkward interaction techniques” [15].

5.1.2 Expert testing

Additionally to evaluating the user’s behaviour, experts can be asked to evaluate the overall usability, user experience and fluidity [14]. That also should happen during the building process to avoid too many adjustments afterwards. There are a lot of rules and usability design guidelines that help creating a system. Experts in the field of interaction design can identify potential problems that users might face during the usage of the system. Afterwards they can help improving them due to their experience and knowledge. Nielsen suggests to include three to five experts to identify the most relevant problems [26]. As this paper was considered looking more at the interaction part of Infovis, not the representational one, we mainly ignored the aesthetics part of our installations. But when it comes to interactivity, the user experiences the behaviour of the whole installation, and that goes closely together with the aesthetics. Our systems involve the ”first-approach-moment” and they are only used voluntarily. So for attracting people to approach them at all they have to have a certain appeal. Interaction designers will also consider the effects of aesthetics and fluidity [11]. Besides getting information out of it, the user also wants to be entertained and enjoy the newest possibilities of technology. ”Users should get a feeling of immersion, first-personness and direct engagement with the objects and the visualizations” [11, p. 4]. If an animation doesn’t react immediately, gets stuck all the time or isn’t visually appealing the chances of people abandoning the installation raise.

5.2 Measuring discoveries

Even though we focused a lot on the approaching process, gaining insight should still be the crucial part of information visualizations. But how can we measure the insight when we actually want the users to independently discover the provided data, rather than asking them to do one specific task? Saraiya et al. defined it in the following way: ”More, valuable, faster, and deeper data findings correspond to more effective visualizations as it suggests users can gain more insight from the data” [29]. Even though their research concerned Bioinformatics Visualizations, they claim that definition to be domain independent. To measure the value and deepness of the findings, they suggest that the experimenters could either use self-reporting and ask the users to rate their own findings or to ask domain experts to rate them [29]. In our cases we think self-reporting could be sufficient, because the data of the installations is mostly preselected and thus it is less likely to discover something completely new (exceptions are possible). So the users could be asked to rate the importance of their findings on a scale, or the evaluators could, due to their expertise in that topic, define the value of a certain finding. The test user thus would achieves a score in the end. Both attempts would yield quantitative results and thus make the evaluation process quicker and less subjective. But until now it is only a suggestion, because there haven't been scientific studies for our specific use cases yet, to prove that method valuable.

5.3 Long-Term studies

The qualitative methods performed during "Into-the-Wild" studies (as mentioned above) seem to yield quite vague results sometimes. The results are strongly dependent on the persons being interviewed and finally the interpretation of the researcher. Also prototype installations often look more like art objects or alien elements especially in public spaces. But as the goal of the visualizations is to inform people, we want to know how much insight they gain through them [29]. And to get an as less distorted result as possible, the user has to be familiar with the system first [30]. Therefore it seems reasonable to observe them over a longer period of time. People then can get used to it and approach them less sceptical. So especially for the evaluation of installations in public spaces that approach should be considered more often. People passing a place regularly, might be in a hurry or just sceptical in the first place, but would maybe approach it another time. However for museums, where people rarely pass by more than once and for news sites, where the topics change really quickly long-term studies are rather inconvenient.

6 Conclusion

There are many great examples for bringing information to people in a digital and interactive way, but still there is a lot of potential for that number to grow. Creating them is connected with a lot of work and the benefits are controversial. For example generating an interactive information visualization for a news website involves editors, designers, programmers and often also a data expert. So it is only profitable for bigger topics that attract many readers and in general also only for bigger companies who can afford experimenting with those tools. But we can see a trend in more and more tools appearing that help building visualizations of data, and are not exclusively accessible for programmers. There is for example "InstantAtlas", "Leaflet", Visual.ly", "Dygraphs", "Google Charts" and many more. Because nowadays also the general "non-expert" is used to computers, the interaction on websites is the most convenient one for everybody and thus creators of interactive graphics can experiment more with new techniques than would be advisable for a museum for example.

For the evaluation there are a lot of established methods available. For the non-expert visualizations considered here it turned out that qualitative methods can bring the most benefit and they can also be combined with quantitative data. But still some improvements could be made when it comes to "Into-the-Wild" studies, which are a useful extension to the isolated lab studies. Even though the lab studies are also important and valuable in many cases, especially for the public installations it is necessary to test them in their actual environment and with untaught people. That again raises some ethical questions, like if it is acceptable to film random people in the public and to invade their privacy by approaching them with a lot of questions. There are some rules that for example the "Human Research Ethics Committee" established and if the researchers are decent and unobtrusive it can be acceptable for all participants.

I intentionally left out the discussion about all the different devices for visualizations and focussed on tabletop displays for museums and desktop computers for news sites. News graphics should ideally also work on mobile devices, but those are mainly still limited in speed and performance so the visualizations can't always be expected to have the same behaviour there as on a desktop computer - also the screen size limits the possibilities of interaction. For public spaces there doesn't seem to be one consistent kind of device yet, as the whole field still is in the phase of development and the rare installations try to use inventive and extraordinary input devices to increase the adventurous experience.

I also didn't go into much detail about how exactly the analysis of the evaluation results is being made. There is a lot of literature for that and especially for the qualitative methods there are some philosophical questions that have to be answered and it is dependent on the interpretation of the particular observer. It may be said that it makes sense to include some more people into the process of analysis to avoid being too influenced by one opinion.

The research also indicated that visualizations in public spaces are not too common yet and differ in many points to those in a museum or on a website. The differences are for example: the target group, the displays, the input devices, the robustness. On a public space one cannot tell a target group at all. Even though museums and websites also are accessible for everybody, it mostly attracts people who are interested in the specific field the museum or website shows. And there is also the problem of maintenance and vandalism but there might be possible alternatives found to build robust and immune displays that can stand all the resistance outside, it surely would be nice to see more public interactive installations in the future.

References

[1] About BELIV 2014 (workshop series focusing on the challenges of evaluation in visualization). `http://beliv.cs.univie.ac.at/about.php`, Accessed: 2014-05-27.

[2] F. Alt, T. Kubitza, D. Bial, F. Zaidan, M. Ortel, B. Zurmaar, T. Lewen, A. S. Shirazi, and A. Schmidt. Digifieds: insights into deploying digital public notice areas in the wild. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pages 165–174. ACM, 2011.

[3] E. W. Anderson. Evaluating scientific visualization using cognitive measures. 2012.

[4] BBC News. Us debt clock runs out of digits. `http://news.bbc.co.uk/2/hi/7660409.stm`, October 2008, Accessed: 2014-06-19.

[5] B. L. Berg and H. Lune. *Qualitative research methods for the social sciences*, volume 5. Pearson Boston, 2004.

[6] J. O. Borchers. A pattern approach to interaction design. *AI & SOCIETY*, 15(4):359–376, 2001.

[7] P. Burnard. A method of analysing interview transcripts in qualitative research. *Nurse education today*, 11(6):461–466, 1991.

[8] D. S. Chung. Profits and perils online news producers' perceptions of interactivity and uses of interactive features. *Convergence: The International Journal of Research into New Media Technologies*, 13(1):43–61, 2007.

[9] D. S. Chung. Interactive features of online newspapers: Identifying patterns and predicting use of engaged readers. *Journal of Computer-Mediated Communication*, 13(3):658–679, 2008.

[10] S. Claes and A. Vande Moere. Street infographics: raising awareness of local issues through a situated urban visualization. In *Proceedings of the 2nd ACM International Symposium on Pervasive Displays*, pages 133–138. ACM, 2013.

[11] N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly. Fluid interaction for information visualization. *Information Visualization*, 10(4):327–340, 2011.

[12] A. Endert and C. North. Interaction junk: user interaction-based evaluation of visual analytic systems. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization*, page 7. ACM, 2012.

[13] F. A. Fabius Steinberger, Marcus Foth. Vote with your feet: Local community polling on urban screens. In *Proceedings of the 3rd ACM International Symposium on Pervasive Displays. Copenhagen, Denmark*. ACM, 2014.

[14] J. L. Gabbard, D. Hix, and J. E. Swan. User-centered design and evaluation of virtual environments. *Computer Graphics and Applications, IEEE*, 19(6):51–59, 1999.

[15] U. Hinrichs, H. Schmidt, and S. Carpendale. Emdialog: Bringing information visualization into the museum. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1181–1188, 2008.

[16] L. E. Holmquist. Evaluating the comprehension of ambient displays. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1545–1545. ACM, 2004.

[17] K. Holmqvist, J. Holsanova, M. Barthelson, and D. Lundqvist. Reading or scanning? a study of newspaper and net paper reading. *Mind*, 2(3):4, 2003.

[18] E. Hornecker and M. Stifter. Learning from interactive museum installations about interaction design for public settings. In *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, pages 135–142. ACM, 2006.

[19] F. Jabr. The reading brain in the digital age: the science of paper versus screens. *Scientific American*, 11, 2013.

[20] B. Jackson, D. Coffey, L. Thorson, D. Schroeder, A. M. Ellingson, D. J. Nuckley, and D. F. Keefe. Toward mixed method evaluations of scientific visualizations and design process as an evaluation tool. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization*, page 4. ACM, 2012.

[21] M. Krauß and M. Bogen. Conveying cultural heritage and legacy with innovative AR-based solutions. `http://www.museumsandtheweb.com/mw2010/papers/krauss/krauss.html`, Accessed: 2014-05-27.

[22] D. Michelis and J. Müller. The audience funnel: Observations of gesture based interaction with multiple large displays in a city center. *Intl. Journal of Human–Computer Interaction*, 27(6):562–579, 2011.

[23] A. V. Moere and D. Hill. Designing for the situated and public visualization of urban data. *Journal of Urban Technology*, 19(2):25–46, 2012.

[24] J. Müller, F. Alt, D. Michelis, and A. Schmidt. Requirements and design space for interactive public displays. In *Proceedings of the international conference on Multimedia*, pages 1285–1294. ACM, 2010.

[25] G. Niederhausern, S. Zhou, E. R. Kang, A. Biderman, and C. Ratti. MIT SENSEable City Lab's EyeStop. `http://senseable.mit.edu/eyestop/`, Accessed: 2014-06-17.

[26] J. Nielsen. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 373–380. ACM, 1992.

[27] A. Peshkin and C. Glesne. Becoming qualitative researchers: an introduction. *NY.: Longman Publising Group*, 1992.

[28] Z. Pousman, J. T. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1145–1152, 2007.

[29] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 11(4):443–456, 2005.

[30] P. Saraiya, C. North, V. Lam, and K. A. Duca. An insight-based longitudinal study of visual analytics. *Visualization and Computer Graphics, IEEE Transactions on*, 12(6):1511–1522, 2006.

[31] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7. ACM, 2006.

[32] J. Somervell, D. S. McCrickard, C. North, and M. Shukla. An evaluation of information visualization in attention-limited environments. In *Proceedings of the symposium on Data Visualisation*, volume 2002. Citeseer, 2002.

[33] Süddeutsche.de and OpenDataCity. Bahn-Pünktlichkeit - Zugmonitor. `http://zugmonitor.sueddeutsche.de`, Accessed: 2014-06-17.

[34] D. S. Tan and M. Czerwinski. Information voyeurism: Social impact of physically large displays on information privacy. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pages 748–749. ACM, 2003.

[35] D. Tewksbury. What do americans really want to know? tracking the behavior of news readers on the internet. *Journal of Communication*, 53(4):694–710, 2003.

[36] R. M. Thomas. *Blending qualitative and quantitative research methods in theses and dissertations*. Corwin Press, 2003.

[37] E. R. Tufte and S. Rifkin. *Visual Explanations: Images and Quantities, Evidence and Narrative*, volume 88. Philadelphia [etc]: Publication and Editorial Office, Dept. of History and Sociology of Science, University of Pennsylvania [etc.], 1997.

[38] L. Van Waes. Thinking aloud as a method for testing the usability of websites: the influence of task variation on the evaluation of hypertext. *Professional Communication, IEEE Transactions on*, 43(3):279–291, 2000.

[39] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1224–1231, 2007.

Memorability in Information Visualization

Stefanie Schreiner

Abstract— Since we went to school, it is important to keep information in someone's mind. Just as it is in Human-Computer-Interaction and especially in Information Visualization. This paper explains how information is perceived and which errors can occur. After the perception information gets in the memory. Therefore, this paper presents the three different types of memory depending how long information remains there. Information visualization is a partition of Human-Computer-Interaction (HCI). Thus, this paper outlines in which areas of general HCI memorability is important as well, besides Information Visualization. Examples are Passwords and the general interaction with devices. In this sectors it is helpful to support the user by presenting its possibilities. To maintain the memorability in information visualization using color is effective. But this is just an example. For different types of information visualizations there are different methods for supporting the memorability. This paper splits information in static and dynamic. It describes whether data is moving or not. Spatializations are one example. It can be effective using 3D landscapes as background to support their memorability. In the dynamic version preserving a mental map has no effect. Memorability can be tested in many ways. Usually, there are two levels. One memorization phase and one for remembering. In this section information is either recalled or recognized. Finally, this paper explains which findings of HCI can be used for Information visualization as well.

Index Terms—Information Visualization, Memorability, Memory, HCI, Graphics, Junk

1 Introduction

Usability is very important in the interaction with computer, apps or environmental displays. It plays a major role in the user experience design. Users should have no problems during their interaction. The ease of use [3] has priority. Usability is defined as "... the capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and user support, to fulfill the specified range of tasks, within the specified range of environmental scenarios [22]." Information visualization is the graphical presentation of data. This is important for presentations in web, newspaper, apps etc. An information visualization should be presented coherently. The user should read information easily. In both areas, memorability is fundamental. Passwords play a major security role in usability. But also in information visualization memorability is important to read data and keep them in mind for presentations et cetera. This paper outlines in which types human memory can be separated. In addition, how to achieve that something remains as long as possible in memory. But before this the paper defines the way of perception, the process before information get in the memory. Further, the role of memorability in general HCI is declared. This is important for the last chapter in which commonalities of HCI and information visualization is discussed. This paper focuses on information visualization which is a partition of HCI. Therefore, the role of memorability in Information Visualization is analyzed in detail. It determines which components support visualizations to keep them in the memory easier. In the next chapter the paper explains how memorability can be tested. Finally, the paper discusses which findings of memorability in HCI can be used for the information visualization.

2 Perception

Before an information gets in someone's memory it has to be perceived to handle and analyze the information. In this chapter levels of the perception and one special phenomenon are explained.

2.1 Preattentive Processing

There are some purposes of objects which are perceived very rapidly without focused attention.These are properties that stand out. It occurs prior to conscious perception. This is important for designing effective visualizations. Treisman [26] examined which visual properties are detected preattentively. She gave users a set of objects and measured the response time of detecting them. The result is that color and shape is perceived preattentively. But not the combination of both. Some other characteristics which are based on color and shape are preattentively mentioned like orientation, length/width, closure, curvature, density/contrast, luminance, intersection, terminators, 3D depth and so on. After the preattentive processing the levels of processing take place.

2.2 Levels of processing

Norman [20] defined three levels of processing, visceral, behavioral and reflective. If an information survives every step it gets in the memory.

The first level, visceral, works fast, automatic and makes quick judgements. It includes the basic perceptual operations of distinguishing objects. Examples are textures and the movement of an object.

The next level is called behavioral. It uses the output of the visceral level and works on it. This level orders objects, perceived in level one, to find a structure of a scene.

The last level, reflective, does not have access to the low-level purposes perceived by the visceral level. It reflects what is happening at the behavioral level, tries to find a meaning and to solve the task.

When an information survives the last level of perception it comes in the memory.

2.3 Change Blindness

Change Blindness [23] is a phenomenon at the perception. It describes the problem of humans not noticing visual changes in a scene. An example is a sudden change in color or movement of an object.

There are some different assumptions why this failure happens:

- The first one i is called overwriting. The new scene replaces the old one in the memory. Thus, there is no basic scene for a comparison and further can not detect a change.
- The next explanation is about the first impression. Here, the first scene is encoded but it fails to encode the second scene. Users have to understand the scene. Then they can detect changes and understand the need of changes in a scene.
- One other declaration is that nothing is stored in memory. There is no need for storing a scene because it is sufficient represented. Further, there is no basis for a comparison.

- *Stefanie Schreiner is studying Media Informatics at the University of Munich, Germany, E-mail: stefanie.schreiner@campus.lmu.de*
- *This research paper was written for the Media Informatics Advanced Seminar 'Hauptseminar Medieninformatik', 2014*

- The next reason is nothing is compared. Users store a mental representation of a scene but there is no comparison made and further no changes can be mentioned.

- The last one is about feature combination. The first and the new scene are combined to a new scene in the memory. There is no possibility of a separation and no comparison possible. [23]

3 Human Memory

After an information survived all levels of perception it comes in the memory. There are three different kind of memories, sensory memory, short-term memory and long-term memory. It depends on how long the remembrance is already in the memory. The first level is the sensory memory. Each memorization which is in the long-term memory has already been in the short-term memory and in the sensory memory before. In this chapter all three kind of memories are explained and how an information comes in there. But first, the whole memory process can be separated in two determinations. There is the implicit and explicit memory, and the declarative and procedural memory. Implicit describes the availability of information without work for recall of the information. An example is, when someone sees a dog on the street, he directly know that this is a dog. Explicit memory determines an intentional work for the recall of information. This describes a process on which humans first have to reflect before knowing the information. Declarative is the knowledge about things, facts and events and procedural describes the process, knowing about how things are done.[29]

3.1 Sensory Memory

The sensory memory keeps an exact representation of physical purposes of sensory stimuli for the duration of a few seconds. There are three different kind of sensory memory. It depends on how a stimuli is noticed.

The iconic memory is the sensory memory for visual stimuli. It stores huge information for a short duration of about one second. An examination of Sperling [24] shows the duration of the sensory memory: Sperling showed users three lines of four letters each. They are presented on a display for a half second. The result is that each user could remember of these letters. This recognition was directly after the presentation. But the longer the duration was after the presentation the less a user could remember of these letters.

The echoic memory stands for acoustical stimuli. It stores like the iconic memory huge information for a short duration. But this short duration is a bit longer as at the iconic memory. It is about four to five seconds. For example, if somebody reads a list of words to someone, every word replaces the earlier one in the memory. Finally, just the last word could be remembered. [29]

The sensory memory for tactile stimuli [14] is called haptic memory. A stimulus is presented and humans store data by touch. Normally, it is used for the interaction with familiar objects. Humans remember the forces which has to be used for the interaction with objects and store surface information of them.

3.2 Short-term Memory

Some information transfer from the sensory memory in the short-term memory. Under Miller [1] the capacity of the short-term memory is of about 7+-2 items. It is called the magical number. The short-term memory stores this information until it is used. Then it is gone. There are two methods for transferring information successfully from the sensory memory in the short-term memory.

The first one is the rehearsal. It is the constant repeating of items in mind, an mnemonic technique. An example is, when someone talks to a colleague and introduces a new name. The colleague has no chance of remembering the name without a brake and repeating the information never gets in the short-term memory.

The second method is called chunking. It is the grouping of a large sequence of numbers or letters in meaningful clusters with history or personal meaning. An example is the sequence, 198919451914. This sequence could be clustered in 1989 1945 1914, with the meanings of fall of the Berlin wall, end of second world war, beginning of the first world war. It is faster and easier getting information in memory and remembering of them as the rehearsal.[29]

3.3 Long-term memory

In the long-term memory information is lifelong available. Information gets from the short-term memory in the long-term memory by repeating the information again and again. There are some improvements of the memorability. First is the elaborate repeating [8]. In this process the information is enriched. While repeating an information make a story of it or choose a visual image which deals with the information. The second one is mnemotechnic. In this improvement linking known information with new information is the principle. One example is called method of loci. Try to memorize a shopping list by going to work. By doing this, each location on the way is associated with an item of the shopping list. For remembering the shopping list going along the way to work mentally.

There is a theory called Levels-of-processing [10]. This indicates that the memory is built on different levels which vary in depth. The deeper an information is processed the higher is the probability of being in the memory. Words or items are coming in there by dealing with them.

When the information is already in the long-term memory there are two possibilities of reproduction. One is called recall, the other one recognition. An example is used to demonstrate the difference. Imagine there is a quiz in the TV. The recall version is that the moderator just asks a question and the candidate has to answer. In contrast, recognition is that the moderator answers a question and provides four additional answer possibilities. That means the candidate gets not only a limit of four answer possibilities but also support for his memorization process. Recognition is easier to reproduce than recall because a reference (answer possibilities) helps finding the right information in the memory. If these reference matches for more than one information in memory, it is called interference.

These references are also needed for the separation of the long-term memory. There is the episodic memory in which personal experiences in life are stored. For this memory type the recall process needs time and space as a reference. While the semantic memory needs no reference for recall. Things like meaning of words or geographical information are stored here.

There is a so-called serial effect of position. If you are learning a list of words and recall them in the same order you will get a good recall performance of the first and the last words. But there is no good recall performance in the middle section.[29]

There are a few hints for making the reproduction easier. One is about the encoding specificity. The recall of information is easier if the reference stimuli during the recall are the same as during the encoding. Godden et al. [13] examined people learning words in two different environments, on the beach or in the water. After the learning process they reproduced those words in their learning environment and in the other. The result was that the performance of recall was 50 percent higher at the same location.

For making the recall easier for someone else it is helpful to know that the human memory is storing things as prototypes. This means that the average of every subjects of a class someone has ever seen in his life forms a typical prototype. This is illustrated by an example. When someone is going on the street, reading a newspaper, watching TV and he sees a dog equal of which class, this dog forms one special image in his mind. Every dog he will further see in his life will deform this prototype of dog he has in his memory. The same effect exists for situations and is called schema.

For the information visualization this could be used for making images or sketchings easier to reproduce. The stronger an image of a thing differs from the mental prototype of the subject the more foreign the thing will appear for users.

The memory is storing things in hierarchies like categories. The higher level of apple is fruit, one level deeper there could be Granny Smith (a special type of an apple). There is a so-called basis level which is an area of the hierarchy. From there is the easiest recall. Seeing in a store the hierarchy, fruit - apple - Granny Smith, most of the humans would say, there is an apple. Therefore, this is the basis level. It is the most specific, less detailed level. We can use this information for visualizations. It is useful not showing too much detailed object visualizations. It is better not to sketch a Granny Smith and the human has to realize, for the deeper meaning of the information, that this is really a Granny Smith. It should be enough to realize that this is an apple. [29]

4 Memorability in Human Computer Interaction

Information visualization is a partition of HCI. Thus, it makes senses to examine the memorability in general HCI first to see which approaches can be transferred to information visualization. This is done in chapter seven. In this chapter focus lies on memorability in HCI. As the field Human-Computer-Interaction suggests this area is based on interactions. Thus, interactions with desktop computers are examined, first. Afterwards, new technologies like mobile and environmental input devices are in the focus.

4.1 Desktop Computer

Desktop computer have general input technologies, mouse and keyboard. There are a few keyboard shortcuts for interacting with computer programs which are hard to remember. The problem is that most of the programs have different ones. One example is Photoshop and Gimp. To select the tool pen at Photoshop the P key has to be pushed on keyboard. To achieve the same result at gimp you have to press key N. For improvement, shortcuts of similar programs could be standardized. Every program in the same field should have the same shortcuts. Another enhancement is that users are able to decide about the assignments of keys themselves. But much better is the user's own arrangement of his typical interactions on the screen. This achieves faster operations as it is possible in Photoshop. At Photoshop, the user can drop functions he often uses directly beneath a picture. This saves deep menu interactions and provides user his selection possibilities as well. This would be even better than shortcuts because recognition is easier than recall (see chapter 3).

Zezschwitz [27] shows how important the layout of programs could be. He investigates the influence of different layouts for an online authentication. This identifies that it is hard using different input programs or technologies even though the user has used it before. Providing one consistent layout for the code input influences the error rate and the memorability of the authentication code.

But not only input technologies are hard to remember. One big point in the HCI are passwords. They are very important because they are the goal of several hacking attacks and there are high security flaws when passwords are weak. But the stronger a password is the more forgettable it is. Thus, strong passwords are difficult to remember. Further, they have to be hard to guess.

Yan et al. [28] examine what a password has to look like to be safe. They prove that a stronger password is harder to memorize. They conduct an empirical investigation of this trade-off. Systems often give advice to choose a secure password. If systems would not give this advice many users choose a weak password. A good password should be long, consisting of capital and small type letters and special characters.

Yan et al. recruited three groups for the study, one control group, a random password group and a pass phrase group.The control group choose the password themselves with a minimum of seven letters and one non letters. The random password group get a sheet with letters from A to Z and numbers from one to nine. They are randomly chosen. Overall, eight characters are noted on a sheet for memorizing. The last group, pass phrase group, get letters from a mnemonic phrase chosen by themselves. This means they choose a whole sentence but just used the capital letters of each word for the password submission.

To prove the force of the passwords they tried to attack each single password. The result was that the random and pass phrase group have the strongest ones. This was the first part of the study. They examined the participants with their passwords for three months. During this term a few users reset their password because they do not remember. Two persons from the control group, one from the random group and three from the pass phrase group. This means that the random group has the best memorable passwords. But no password technique was perfect.

One possibility of improving the trade-off can be pictures because they are easier to remember than text [18].

Davis et al. [11] investigated this property and examined another way of determining and remembering a password. They investigated the probability of an attack for a graphic password when the attacker has access to demographic information of a participant.

The study provides a four times 3x3 grid of difficult images of faces. In each grid the user choose one face which presents the password face in this grid. Davis et al. identified that chosen faces depend on race of the participant and the attractiveness of the face. Generally, faces have a high degree of memorability.

This study is taking place during a whole term at university. They compared a faces password to a story password. Story passwords have the same set as faces passwords but with images like landscapes, cars et cetera on it. The results are that faces password had more correct logins. Further, the memorability of faces passwords was most successful since last login during the term.

4.2 Mobile and environmental input device

It seems that the modern tendency is using more mobile and environmental input devices. But there is the problem of remembering a lot of input alternatives for different devices because there is no mouse and keyboard anymore which everyone knows to interact with. For environmental and mobile devices other input technologies are needed. New input standards has to be learned. Normally, for mobile input devices direct touch on display is used. Noel et al. [19] showed that new interaction techniques are difficult to learn and to remember for mobile devices. They investigated the memorability of mobile in-car navigation.To examine this they conducted a study of paired users in a car interacting with a navigation system. After one month there is a second study with the same experiment as in the first one. It was shown that participants has the same problems as in the first study. They remembered about the problem but not how they solved it in the first session.

The same counts for the input of environmental devices. Bowman et al. [9] shows a good example of a new input methodology for users to support memorability on its best. They developed so-called Pinch gloves for environmental interaction. These gloves report contact between two or three fingers as input for the system. They developed three different interaction techniques. For this paper just one technique is explained to get insight.

Bowman et al. developed a Menu system. The first thought was doing natural gestures like connecting thumb and the second finger which normally in diver language means ok. But not in all cultures this meaning is the same and in general there are very less natural finger connections. Then they decide to connect fingers with thumbs to have eight different interactions. On the non dominant hand they defined to be the top level menu labels, and on the dominant hand are the second level labels. Because this is hard to remember they decide to present fingers with labels on the screen which makes recognition again easier than recall for memorization.

5 Memorability in Information Visualization

Information visualization is the base of good and effective presentation of data. Users should understand the information easily and the visualization should be clear and not overcrowded. In general, we specialize in memorability in this paper. In this chapter results of different research paper which investigate the memorability of information visualization are presented. They are ordered by the kind of information, static and dynamic information. Static information is data which do not move in the presentation while dynamic information does.

5.1 Static Information

Images are a little bit difficult to assign because somehow they belong to the general HCI chapter as well. It depends on what you see on the image. The study of Isola et al. [15] investigates the memorability of common images. They separated images in pattern for investigating which purpose is better memorable. There are a few different results for the study of Isola et al. Thus, memorability is supported by red and green coloring to Colors blue to purple achieve the opposite. The mean saturation and the value have weaker correlations with memorability. The different object regions do not correlate with memorability. People, interiors, foregrounds and human-scale objects support memorability on images. Whereas exteriors, wide angle vistas, backgrounds and natural scenes correlates negatively with memorability.

Information visualization can use such common images. One example for this is the work of Bateman et al. [5]. They use image components by adding sketches to the information. Bateman et al. examined the memorability of sketched information by Nigel Holmes. Holmes is an artist who designs graphics. He pictures information by using junk. Junks are illustrations which are not essential for the data. In figure 1) the monster is not neccessary for understanding the different costs of elections. Therefore, the monster is junk. Borkin et al. showed people different information images. Once the data is presented as a plain version, once as a junk version *(see figure 1)*. For the immediate recall there are no differences between plain and Holmes graphics. But for the long-term recall (a few weeks later) there is a significant difference for memorability. More charts and details are remembered than in the plain version. This paper is very disputed. Edward Tufte[2] criticizes chartjunks like Holmes graphics. He thinks that this tells the user nothing more. It is not informative and presents non-data which is redundant. Those chartjunks are just non-informative and frivolous. The paper of Bateman et al. uses just Holmes graphics which is a popular artist. But there are many other chartjunk creator which do not support an information. Graphics are always better memorable than plain graphs. The graphs used are very simple, too simple for a comparison with graphics. Nomally, they are not just as plain as they are used by Bateman. They are very ugly with no style principles. Normally, they are more colorful and not just grey. Color is more memorable at all. But Bateman used plain graphs to support their test results.[21]

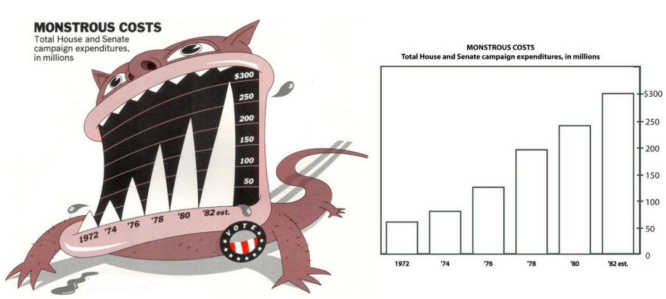


Fig. 1. Different presentations of the same information. The left one is sketched by Holmes, the right one is the plain version[5].

Borkin et al. [7] also used images for their study. They showed test persons images with different kind of data presentations with variation of color, form and style. If the users see one image the second time in a sequence they should press a key. The results are that visualizations are less memorable than natural scenes. Furthermore, minimal visual density of visualizations is less memorable than a high one. Pictorial, grid/matrix, trees, networks and diagrams have significantly higher memorability scores than circles, areas, points, bars, and lines. This is because novel and unexpected visualizations can be better remembered than the visualization with limited variability. Visualizations are more memorable by using color, including a low data-ink ratio, the inclusion of human recognizable objects and pictograms and cartoons of recognizable images.

Besides visualizing information by images many other types exist. In some areas the connection of data is important. One example are networks. Marriott et al. [17] investigated the representation of network diagrams. They are exclusively presented by connected nodes and lines. They want to test the effect of their different alignments on the recall. The result is that the node-alignment and parallel lines of graphs are not reproducible for user but aid recalling. Features like symmetry, collinearity and orthogonality strongly supports memorizing graphs.

But there are more possibilities than just changing the arrangement of information visualizations. Lam et al. [16] examined the influence of manipulating information. They measured the effect of scaling, rotation, rectangular fisheye, and polar fisheye transformations on visual memory. The diagrams consist again just of connected dots and lines. They also want to investigate if an additional background grid supports visual memory. The result is that there is no memorability effect for scaling and fish-eye with or without grid. With Polar fish eye people can not exactly remember if they have seen the visualization before or not. For rotation the response time is better with no grid. With a rotation down to 60 degree there is a decrease of response time with and without grid. The response time determines the period in which the test person answers if he has seen a visualization before.

Another kind of information presentation are spatializations. In this area position of objects are important. Tory et al. [25] examined the memorability of spatializations. They compared the memorability by using dots and landscapes. The study has two results. For having more dots in a spatialization people can say easier that they have not seen a spatialization before. In general, dots and 3D landscapes combined are most accurate *(see figure 2)*. In this case people can say confidently that they have or have not seen a visualization before.

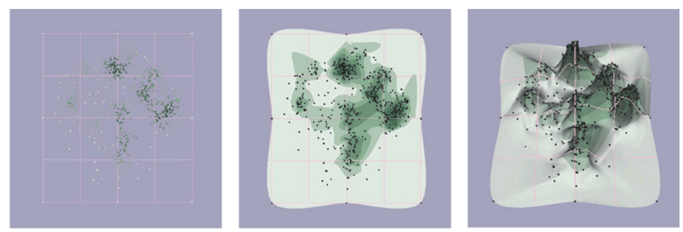


Fig. 2. Left shows just dots, the middle image shows dots with 2D landscape and the right image shows information with dots and a 3D landscape with 500 points[25].

5.2 Dynamic information

As in the static information chapter the connection of information can be visualized by links and dots. Archambault et al. [4] examined the memorability of those graphs but in dynamic version. The challenge in this test is remembering of changing nodes and links over time. They want to show that this depends on a mental map preserving. This is simulated by keeping the same node in the same area of the plane just adding new nodes*(see figure 3)*. But the result was that preserving the mental map has no effect on task performance.

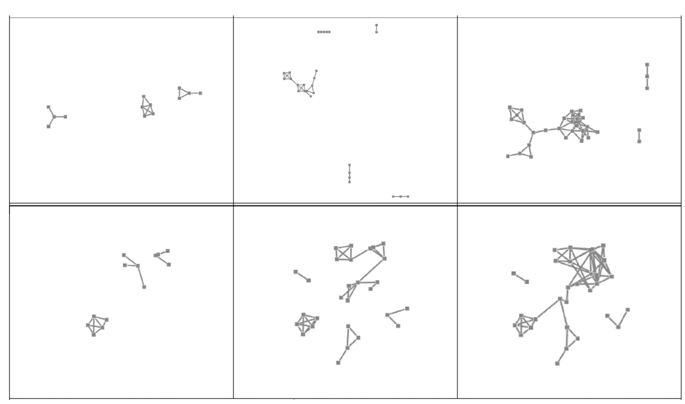


Fig. 3. The upper image cue shows the graph with no preserving of the mental map. The lower one shows the graph with mental map preserving[4] .

But preserving a mental map is a good beginning for supporting the memorization of dynamic information. Bederson et al.[6] examined the influence of animation to build a mental map. They want to test the effectiveness of animation on subjects ability to build mental maps. They tested online family trees in two versions. Once the relationship to another person of the tree is animated, once it is static and just jumps to the selected other person in the tree. Their hypothesis is that animation improves subjects ability to navigate through an information space, recall the information and finally to reconstruct it. The result is users had more recall errors in the non animated tree. If they did the animation task second the performance was better. If users did it first there was no difference.

There is also an dynamic version of image information. Dinh et al. [12] investigated different sensory inputs on memory in a virtual environment. This environment is a walk-through in a virtual flat. The users can decide by themselves when and where to go. As one might expect the virtual environment belongs to the general HCI chapter. But in this paper they examined tactile, olfactory, audio and visual sensory cues additional to the image information. That means it is more about which sensual information is realized better for humans and which information stays in the memory. The result is that there were no special differences. But for the audio situation more users remembered that there was a toilet. When they passed the bathroom in the test they hear a toilet flush. For the olfactory and tactile cues the recall was higher for object locations.

6 How to test memorability

There are different ways for testing memorability. In this chapter testing processes from the mentioned paper of the previous chapter are presented.

6.1 Test in a sequence

The first procedure to test memorability consists of one level. Users are shown a sequence of images. If users think to view an image for the second time they press a key. This sequence consists of target images and filler images. Target images consist of special purposes. These purposes are chosen by the authors. They think that these purposes support the memorability. Each image is shown for a specific amount of time. [15, 7]

One difficulty in this procedure is the choice of the right amount of time to show a picture. If it is too short, than, there is only the chance of viewing the layout of an image. The second difficulty is that in this short second not only the memorization phase takes place, but also the reproduction. The first step of the participants is that they need to consider whether they have already seen the picture. After that is done memorization takes place. That is too much for such a short period. The third difficulty determines the human memory. In one second information just can get in the sensory memory (see chapter 3.1). There is no further time for dealing with information. Therefore, layout information remains in sensory memory and data information is not perceived. Information is getting lost immediately.

This method is useful when the layout wants to be investigated. It focuses on what stands out the layout. The data presentation is accidental.

6.2 Two levels with free reproduction

Another procedure consists of two levels, one memorization phase and afterwards a free reproduction phase. In the memorization phase users have enough time to bring content in the memory. Two examples are clicking through two different versions of family trees [6] or memorizing different network graphs [17]. After this, a free reproduction is taking place. This can mean sorting images of family members to names and hierarchies [6] or painting network graphs on a tablet like users might have seen them [17]. Both is measured by error rate and accuracy.

One difficulty of this procedure could be that users have too much time for memorization. They do not know when to stop with memorization. Finally, they memorized the whole data and no standing outs can be measured.

This method can be used when the content is in focus. Participants have enough time for interpreting data and memorizing them. Furthermore, the memorability of information can be tested.

6.3 Three levels to get familiar

The next procedure consists of three levels. The first level is to get familiar with the environment. After this is done participants get in the test environment. At Dinh et al. [12] this means that users go in the virtual environment. They have as time as they they want. They walk around in a virtual flat with different stimulations. The last level starts when users are done with level two. They have to answer questions about what has been noticed. The questions can be about presence, spatial layout or the object location in this case.

One difficulty is that by answering fixed questions at reproduction phase findings can be limited. Results which are covered by questions can come out. But nothing more. On the opposite this can be useful if the tester really wants just a special result.

Therefore, this test can be used when tester want to test fixed components to prove selected results.

6.4 Two levels including a break

Another procedure consists of two phases again. One is for memorization, the other one for reproduction. At the memorization phase, different pictures are shown in a cue. By watching one image participants were ask a few questions. At Bateman et al. [5] They showed participants several charts. Each chart is available in two versions. One is plain and the other one is sketched by Holmes. One participant group saw the Holmes slide show, the other one the plain slide show. The special feature of this method is that the participants know nothing about the second phase. The second phase of this method is the recall. The group of participants is split. One group did an immediate recall with a short gaming break between. The other group did a recall with a break of two to three weeks. The recall phase is a free telling procedure. Participants should try to mention as many charts as they are able to remember. Afterwards they try to explain the charts as detailed as they could based on the four questions of the first phase.

The difficulty is the influenceability of this method. By asking questions during memorization participant's attention is specifically drawn to something. Questions can cause that something is perceived, which would otherwise be lost. That means the result can be different as in a neutral situation.

This procedure is optimal when not only short-term recall is tested but also long-term recall. When the difference between both is important this is the right method to choose.

6.5 Sequence with two levels

The last method has a memorization and a reproduction phase, too. In the memorization phase participants view graphs with different components/stimuli in a cue. Users have enough time to view graphs. In the reproduction, they are shown a graph cue again. Participants have to say if they have seen a graph during memorization or not. Usually, half of graphs are from memorization, half are new in the reproduction cue. [4, 16, 25]

This procedure can be used when more complex content wants to be tested. Participants have enough time to memorize graphs. By using a recognition phase in the end user have a simplified reproduction phase.

Finally, each paper uses either recall or recognition as type of reproduction (see chapter 3) *(see figure 4)*. Usually, the mentioned methods use two main phases. One for the memorization and second for the reproduction of the content. Just Borkin et al. and Isola et al. used one phase in which both, memorization and reproduction take place. For getting an overview which results from chapter five used which testing method*(see figure 5)*.

7 Discussion

Information Visualization is a partition of HCI. Further, both have some commonalities. In HCI, the focus lies on interaction. This has several advantages for memorization. When a person interacts with a system he always does this by using this system. He deals with it

	type of remembering	
	recognition	recall
Borkin et al.	x	
Isola et al.	x	
Bederson et al.		x
Marriott et al.		x
Dinh et al.		x
Bateman et al.		x
Tory et al.	x	
Lam et al.	x	
Archambault et al.	x	

Fig. 4. Comparison of remembering types of mentioned paper.

	first contact phase		remembering phase			one sequence
	memo phase	test environment	free reproduction	questions	graph/image sequence	press key
Borkin et al.						x
Isola et al.						x
Bederson et al.	x		x			
Marriott et al.	x		x			
Dinh et al.		x		x		
Bateman et al.	x		x			
Tory et al.	x				x	
Lam et al.	x				x	
Archambault et al.	x				x	

Fig. 5. Comparison of remembering types of mentioned paper.

by exploring possibilities. By doing this, the information is repeated in memory and stays in memory. That is one reason why interactive information remains better in someone's mind.

But there exists also a second advantage. Interactions are system dependant. That means each system needs its own interaction technique. Using the same environment again and again for one explicit task helps information to stay in memory. The reference stimulus during the recall has to be the same (see chapter 3.3). In this case the reference stimulus is the system. By interacting with the same system again it provides the reference for remembering and an informative basis. In general, HCI uses recognition.

These both advantages can be used for information visualization. But just in case an interaction is possible with the visualization. By filtering an information visualization a system provides interactions. In this situation findings of the HCI transfers interaction techniques in the memory. New technologies are possible as well like gloves (see chapter4.2). In such an interaction advantages of the HCI are essential.

But usually, information visualizations are visual presentations of data in newspapers or scientific paper. There is no interaction possible. That introduces new requirements. Visualized data sets are huge and specific. And it is like learning facts for school. Much scientific information has to get in memory in detail. One big question is when do people need such information in their memory. They need it for presentations, for exams, for their work or just telling other people about. In this case it differs from HCI. There is no common system that provides interaction possibilities or reference stimuli. Especially when people hold presentations or write an exam about it. There are no references in environment. Usually that means in information visualization is no recognition but recall. People are left on their own for memorization. This fact makes it even more important supporting the visualizations for an easier memorization. This is the only way supporting static information. As the studies in this paper showed every sense can be used for achieving this, even the visual one is most important.

One huge problem is still dynamic information. In this case, it is most difficult to remember. As studies showed there are two differences in dynamic information. In one situation the information is moving[4]. In the other case the user is moving[6, 12]. The second situation is easier because the user decide which part of information to see. Further, user deal with information and support memorability. The first case is just complicated. User do not know how the information changes. They first have to see which information changed and how. They have no time to focus on how the information looks like and what it means. For memorization time is too short. The next change happenes now. The same procedure again with no or less time for interpreting and memorization. In this area there is still much work to do for supporting memorability.

8 Conclusion

As the previous chapter showed there already exist a few principles which support the memorization of information visualizations. In general, it is not easy to bring something in the memory, especially in the long-term memory. In HCI references help user to remember about interaction techniques. In information visualization they do not exist. Therefore, it is more important to design visualizations which support user by remembrance. To achieve this, it has to be started at perception. Important things has to be emphasized. Using color is a good beginning to get an improvement. Which properties to use depends on the type of information. But in general, every purpose that can support memorability has to be used economically. Otherwise, the converse is achieved. Especially, in the field of moving information there is a lot of work to do. In future, some purposes which support memorization have to be found in that area. To test whether a purpose supports memorability, a suitable test method has to be chosen. Which one to take is dependant on what is tested.

References

[1] *The Psychological Review*. George A. Miller, 1956.

[2] *The visual Display of Quantitative Information*. PR Graves-Morris, 1983.

[3] Usability in e-learning, 2005.

[4] D. Archambault and H. Purchase. The mental map and memorability in dynamic graphs.

[5] S. Bateman, R. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk? the effects of visual embellishment on comprehension and memorability of charts, 2010.

[6] B. Bederson and A. Boltman. Does animation help users build mental maps of spatial information? IEEE Proceedings of Information Visualization Symposium, 1999.

[7] M. Borkin, A. Vo, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. Makes a visualization memorable? IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, December 2013.

[8] G. Bower. Mental imagery and associative lerning. Cognition in Learning and Memory, 1972.

[9] D. A. Bowman, C. A. Wingrave, J. M. Campbell, V. Q. Ly, and C. J. Rhoton. *Novel Uses of Pinch GlovesY for Virtual Environment Interaction Techniques*. Springer, 2002.

[10] F. Craik and R. Lockhart. Levels of processing: A framework for memory research. Journal of Verbal Learning Verbal Behavior, 1972.

[11] D. Davis, F. Monrose, and M. K. Reiter. On user choice in graphical password schemes. USENIX Security Symposium, 2014.

[12] H. Dinh, N. Walker, C. Song, A. Kobayashi, and L. Hodges. Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments, 1999.

[13] Godden and Baddeley. Context-dependent memory in two natural environments on land and underwater. British Journal of Psychologie, 1975.

[14] A. Gordon, G. Westling, K. Cole, and R. Johansson. Memory representations underlying motor commands used during manipulation of common and novel objects. Journal of Neurophysiology, 1993.

[15] P. Isola, A. Torralba, and A. Oliva. What makes an image memorable? IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[16] H. Lam, R. Rensink, and T. Munzner. Effects of 2d geometric transformations on visual memory. Proceedings of the 3rd symposium on Applied perception in graphics and visualization, 2006.

[17] K. Marriott, H. Purchase, M. Wybrow, and C. Goncu. Memorability of visual features in network diagrams. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, December 2012.

[18] D. L. Nelson, U. S. Reed, and J. R. Walling. Human learning and memory. Journal of Experimental Psychology, 1977.

[19] E. Noel, B. Nonnecke, and L. Trick. Evaluating first-time and infrequent use of in-car navigation devices. Third International Driving Symposium on Human Factors in Driver Assessment, 2005.

[20] D. Norman. *Emotional design: why we love (or hate) everyday things*. Basic Books, 2004.

[21] S.Few and P. Edge. The chartjunk debate the chartjunk debate a close examination of recent findings. Visual Business Intelligence Newsletter, 2011.

[22] B. Shackel and S. J. Richardson. *Human Factors for Informatics Usability*. Cambridge University Press, 1991.

[23] D. Simons. Current approaches to change blindness, 2000.

[24] G. Sperling. The information available in brief visual presentations, 1960.
[25] M. Tory, C. Swindells, and R. Dreezer. Comparing dot and landscape spatializations for visual memory differences. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 2009.
[26] A. Treisman and G. Gelade. A feature-integration theory of attention.
[27] E. von Zezschwitz. An evaluation of the influence of external factors on authentication performance and memorability. Master's thesis, Ludwig Maximilans University Munich, 2010.
[28] F. Yan. Password memorability and security: Empirical results. *IEEE Security Privacy*, 2004.
[29] P. G. Zimbardo and R. J. Gerrig. *Psychologie*. Pearson Studium, 2004.

Please Select Your Password!
On Increasing the Practical Password Space

Ismail Kosan

Abstract— Users of web services are often confronted with choosing a username and an alphanumerical password for authentication. The security of their accounts depends strongly on the security of the chosen passwords. Nonetheless most users tend to choose practical passwords which are easy to remember and unfortunately also easy to guess. Recent password cracking studies have shown that a considerable amount of real life passwords could be cracked very easily (up to 50%). Attackers make intensive use of predictable user behavior and further human factors to achieve such high success rates. This makes clear that strategies are needed which will guide users towards selecting stronger passwords. This paper presents such strategies which aim to increase the practical password space. First, traditional methods are presented, which either enforce or recommend users to create stronger passwords. Then novel approaches are introduced, which give users persuasive feedback or make use of the input interface. The impact of these methods on password selection shows that they can indeed guide users to create stronger passwords.

Index Terms—Passwords, Security, Usability, Authentication, Selection, Human Factors

1 Introduction

More and more people use online services to create and store their private and personal data. They use them to communicate with other people, to work and do business, for shopping and entertainment, to manage bank accounts and much more. To ensure that only the authorized person has access to these important services and data, secure ways of authentication are needed. In general there are three different types of authentication mechanisms [36]. Something the user knows (knowledge-based), for example a password or a PIN, something the user has (token-based), for example smart cards or physical keys, and something the user is (biometric-based), for example fingerprints.

For now, the most commonly used authentication mechanism in the digital world is the knowledge-based authentication, especially the selection of a user name in combination with an alphanumerical password [27]. Alphanumerical or text-based passwords have been well established since the early use of computers. They are simple and effective to use and implement, can easily be changed, require no special hardware and are based on secret knowledge of their users.

It is very important that users create secure passwords. If an attacker could guess a user's password, he could compromise sensitive accounts like online banking or e-mail, monitor and manipulate communication, steal private data from cloud services or do other malicious actions.

On the one hand passwords have to be secure to make it unattainable for an attacker to crack or guess them. On the other hand passwords have to be usable, in particular easy to remember. For password selection it is challenging to ensure both. There is a tradeoff between security and usability [11]. Random generated passwords are secure and hard to guess but difficult for users to remember, while user-chosen passwords are easy to remember but they also tend to be highly predictable [11, 24, 57]. Studies about leaked real life password databases have confirmed that most of the users select simple passwords, for example "123456", which are easy to remember but unfortunately also easy to guess [4, 16, 59]. Nowadays an average web user has to handle about seven passwords frequently and has a total of 25 for his accounts [16]. It would be very hard for a user to remember 25 different passwords if all of them had been generated randomly, such as "XE%#6cQ{9*Ze". Users need practical solutions to develop secure but also usable passwords.

- *Ismail Kosan is studying Media Informatics at the University of Munich, Germany, E-mail: i.kosan@campus.lmu.de*
- *This research paper was written for the Media Informatics Advanced Seminar 'Special Aspects of Usability', 2014*

The goal of this paper is to identify and discuss methods which guide users to select stronger passwords and increase the practical password space.

2 Password Security

The security of alphanumerical password authentication depends strongly on the strength of the passwords used [8]. Theoretically the strength can be measured by the difficulty to determine or guess a password. Alphanumerical passwords are strings of a selected length that consist of a set of printable characters. If the set of all printable characters (except space) from the ASCII table is taken, there would be an alphabet of 94 available characters for creating a password. For example a string of 8 characters with a used alphabet of 94 characters would have a space of $94^8 \approx 6.1 * 10^{15} \approx 2^{52}$ different possible passwords. In general the total number of all selectable passwords, the theoretical password space, can be calculated by the formula a^l (a is the length of the used alphabet, l is the length of the password). This implicates that the password strength strongly depends on the length and the amount of selectable characters. The longer the password is and the bigger the alphabet is, the more guesses are needed by an attacker to determine the password. For example an attacker would need on average about $2.4 * 10^{23}$ guesses to determine a 12 character long password with an alphabet of 94 characters. If he could try 3 trillion guesses per second it would take about 2515 years to guess the password. Under this assumption this type of password would be very hard to guess and therefore very strong and secure.

But in the context of user-chosen passwords there is an important problem with this approach for measuring the password strength only dependent on length and character set. This method assumes that all characters have the same probability distribution and that passwords are generated randomly. As users commonly don't use all of the possible characters and use some characters more often than others, the characters probability is not equally distributed, for example the symbol "!" is used significantly more often than the symbol ">" [48]. An intelligent attacker would try the more probable combinations first. With this aspect the password strength additionally depends on the probability distribution of the characters. Therefore there are other approaches to measure the strength of passwords taking this factor into account.

2.1 Entropy

One common approach to measure the strength of passwords is to use the concept of information entropy. Entropy has its first formalized roots in Shannon's mathematical theory of communication [44]. It is an approach to measure the amount of information that is unknown due to random variables. Shannon used this method to explore the uncertainty of letters and words in the English written language and developed a way to estimate their probability distribution, for example in printed English it is more likely that an "w" is followed by an "h" than another "w" [43].

In the context of passwords this method can be used to measure how difficult it is to predict the value of a character [8]. The predictability of the password can be calculated by summing all entropy values derived from each component of a password. For example the placement, number and content of each character can separately contribute to the sum of the entropy of the whole password. The idea is, that in an optimal attack an attacker would try the highest probability passwords first and entropy could measure the expected number of tries until success. Therefore the more entropy a password has, the more difficult it is to guess. Entropy is usually given in bits. For example a password with the entropy of 18 bits would have a strength of $2^{18} = 262144$ possibilities to be guessed.

The US National Institute of Standards and Technology (NIST) uses this method in their published document, the Electronic Authentication Guide SP-800-63 [8], to make recommendations for measuring the strength of human generated passwords. NIST guidelines suggest that a password entropy can be measured by the sum of each of the following characteristics. The first character gets 4 bits of entropy. 2 bits are given for each of the next 7 characters. The 9th to 20th character has 1.5 bits each, the 21st and above ones have each 1 bit of entropy. A bonus of 6 bits is added if uppercase letters, numbers and symbols are used. Further 6 bits are added if the password is not contained in a dictionary. With this measuring a password like "password" would have 18 bits of entropy, "P@ssword1" would have an entropy of 26 bits.

Password security in general does not only depend on the strength of the password, there are several other factors. For example the technical implementation or social attacks can affect password security [36]. If passwords are compromised by keyloggers, phishing or other malicious manipulation then even the strongest password is useless. In this paper such factors for password security will not be considered. The focus lies on the security regarding the alphanumerical password space.

2.2 Theoretical vs. Practical Password Space

As mentioned above, alphanumerical passwords have a theoretical password space. It is based on the point that passwords are generated randomly with an equal probability distribution for each character and that they follow no known or predictable pattern. This theoretical password space would be large enough to prevent an attacker to guess passwords in a reasonable amount of time.

However, usually users do not choose their passwords randomly, their choices can be predictable [2, 54]. Human factors transform the theoretical password space into a practical password space, the number of passwords that are likely to be chosen by real world users.

Users face the challenge to remember multiple different passwords for several applications [16]. Therefore many users choose easy memorable passwords, which are also often easy to guess [1, 38, 61]. Passwords which are hard to guess have shown to be more difficult to remember [61]. In a study of Adams et al. [2] half of all users reported that they use their own strategy for creating multiple passwords by using one common element which is linked to all related passwords, so they would be easy memorable for them. For example having to change the password frequently led users select passwords, which follow predictable patterns, such as "tom1", "tom2", "tom3" [1, 38]. Zviran et al. [61] have found that most of the user-chosen passwords are made up of characteristics of personal details meaningful to the user and are relatively short (4-7 characters). 80% of their participants claimed to use only alphabetic characters. Only 13% used additional numbers and less than 1% used symbols. A recent study by Veras et al. [54] about real life leaked passwords have shown that many chosen passwords contain concepts relating to love, sexual terms, profanity, animals, food, and money. A study of Das et al. [12] revealed that about 43% of website users reuse their password across several websites, either unmodified or with only small modifications. Von Zezschwitz et al. [55] have studied the evolution of password changes of users. They reported that, despite users use significantly more characters to build stronger passwords, most of them also rely on their early weaker ones. It has been shown that the security of users' passwords across different websites depends on the weakest one chosen [19]. An attacker could guess the weakest of all passwords and modify the password to guess more passwords and get access to more sensitive data of other web services. In a large scale study of users web password habits, Florencio et al. [16] have shown, that users have on average seven passwords, each of which is shared across 3.9 different sites. Each user has about 25 accounts to manage and types an average of eight passwords a day. Furthermore most of the users used lowercase letters only in their passwords, unless they were not forced to do otherwise. The average strength of user chosen passwords was 40.54 bits. If this is compared with a randomly generated password with the length of 8 characters using an alphabet of 94 characters (which would correspond about 52 bit entropy), then the average strength of user chosen passwords would only make up about 0,025% ($2^{40}/2^{52}$) of the randomly generated passwords. By analyzing leaked password databases Dell'Amico et al. [14] showed that the average password was only 8 characters long. Further studies by Riley [38] have shown that about 55% of users used personally meaningful words, such as words of children, pets or street names, while about 50% of users used personally meaningful numbers, such as birth dates or telephone numbers. Further human password practices have been revealed by Brown et al. [5]. They reported that two thirds of users created their password with information about personal characteristics, most of them regarding to relatives, friends or lovers. About half of their studied password constructions were based upon proper names and birth days. People mostly select passwords this way, because they have to be easy to remember for them and security often plays a less important factor [59]. This fact can also be observed when it comes to aspects of maximum security. For almost 20 years the secret start code to prevent unauthorized people to launch US nuclear missiles during the Cold War was "00000000" [18].

If humans did not have memory limitations then passwords could have maximum entropy. However, the human memory is temporally limited for sequences of items [59]. It has a short-term capacity of around 7 +/- 2 items and if they are remembered, those items have to be familiar chunks such as words or familiar symbols [59].

Human factors and predictable user behavior result in a password space which is much smaller than the theoretical password space. Obviously the weakness in password authentication lies in the choices the users make. Taking the human factors into account it can be said that a good password would be one which is easy to be remembered, but yet difficult to guess. To achieve this goal there have to be strategies to increase the practical password space, so that it would be too large for an attacker to exhaustively guess passwords in a reasonable amount of time, but at the same time offer users the ability to create usable/memorable passwords on their own.

2.3 Password Cracking

An attacker can benefit from the insights about users' typical password choices, to develop intelligent algorithms for password cracking. In the following different ways of password guessing and studies about real life password databases are presented.

2.3.1 Brute Force Attacks

Brute-force attacks address the theoretical password space. Brute-force attacks are search processes, in which all possible character combinations are built and tried to guess the correct password. These attacks can be successful if an attacker has enough computational power and the password space is relatively small, so that it would not take too much time to crack the password. As an example, a password with an entropy of 40 bit would have about 1,1 trillion possible combinations. Assuming an attacker, who could guess about 3 billion passwords per second, would need on average about 3 minutes to guess the correct password. To ensure appropriate protection against brute-force attacks, the password space has to be large enough, so trying all possible combinations would need too much computational time. According to the same assumptions as above, it would take about 98 years to guess a password with 64 bit entropy.

2.3.2 Dictionary Attacks

Dictionary attacks address the practical password space. Instead of trying all possible character combinations to guess the password, dictionary attacks use predefined lists of words which are likely to be selected as real passwords. These lists can contain whole dictionaries of spoken languages, names of famous stars, personal information about people, such as nicknames and birth dates, or any other meaningful word. Usually they contain words, which are known to be selected frequently, such as "password", "123456" or "letmein" (from the Top 500 dictionary of worst passwords of all time [35]). Theoretically an attacker could put any word into a dictionary, of which he thinks that it could be the correct password. This is very interesting in context of social engineering, because if people use personal information within their passwords, such as ID numbers or family nicknames, an attacker could put and combine these information into his dictionary.

Dictionaries can also be extended by generating new candidate passwords with leeting or mangling rules. Mangling rules apply different transformations to passwords such as prefixing or suffixing with digits or symbols, capitalizing or reversing a password and combinations of them. Leeting rules are based on the fact that there are similar visual equivalences between letters, digits and symbols. This means that characters are transformed into similar characters, such as transforming an "a" to an "@" or a "s" to a "$". With these modifications the password "password" could be transformed into "Password", "drowssaP", "password1", "p@ssw0rd", "P@$$w0rd123" and so on. These transformed dictionaries are often used to address password policies, which force users to select passwords of mixed characters. Many users adopt these simple mangling or leeting techniques to their simple passwords to bypass policy restrictions [14, 48, 52].

In the 1990's password cracking tools which automatically performed dictionary attacks emerged, for example the popular tool "John the Ripper" [31]. Studies have shown that these cracking tools can achieve far more success in guessing passwords than the NIST entropy predicts [14]. Nowadays there are various dictionaries available for such tools, consisting of millions of words. For example the OpenWall Mangled Wordlist consists of about 40 million words and the dictionary of leaked passwords from the RockYou website has a size of 32 million [32, 42]. Using such tools and dictionaries has shown that passwords can be guessed in a considerable amount of cases. In studies of Klein [25] and Spafford [49] 22-24% of passwords were guessed by using dictionaries containing lists such as of names, sport teams and movies. Dell'Amico [14] et al. could guess about 48% of the passwords from the leaked MySpace password database using a dictionary consisting of about 148 million entries. In recent studies 20%-50% of leaked password databases could be cracked by using dictionary sizes in the range of $2^{20} - 2^{30}$ [4, 57].

Researchers have recently developed even more successful guessing algorithms which benefit from the non-uniformity in character sequences and the predictability of password structures. Narayanan and Shmatikov [30] use a password cracking algorithm which is based on Markov chains. This method makes use of the fact that the distribution of character sequences within passwords is not equally balanced. By training Markov chains on known password dictionaries they estimated the distribution of character sequences and used them to generate a list of passwords which was significantly more effective than random guessing and the cracking tool "John the Ripper". Weir et al. [58] developed a way to crack passwords by using probabilistic context-free grammar (PCFG). This approach is based on the observation that passwords have predictable structures. Passwords can be divided into strings of letters, digits and symbols. For example the password "Password123" would have the PCFG structure "U1L7D3", which represents an uppercase letter followed by 7 lowercase letters followed by 3 digits. This information can be useful for an attacker to calculate the occurrences of each structure from a training set and develop a list of the most probable structures. Using this method Weir et al. could guess 25% of passwords which are more than 7 characters long with a dictionary of the size 50000. With this approach they observed that the NIST entropy can only be a rough approximation for password strength. They suggest that it would be a more realistic approach to determine password strength by calculating the difficulty to guess passwords with modern cracking algorithms.

It is very important to keep in mind that the success and effort of password cracking depends on whether passwords can be attacked online or offline. In an online attack there may only be few passwords which could be tested automatically until the service will rate limit the attempts or requires the input of a *CAPTCHA* (a test, where the user has to type the characters of a distorted image, to determine that the user is a human and not a computer). Whereas in an offline attack the attacker would only be restricted by his computational power. Since password databases like the RockYou or the MySpace databases have already been hacked, there is no guarantee that this could not happen or has already happened undetected to other password databases. Quite recently the password database of eBay has been hacked which forces about 145 million users to change their passwords [34]. This recent incident also confirms the assumption that offline password guessing should be considered as the worst case scenario for password cracking.

3 Traditional Methods

Adams et al. [2] have found that users fail to understand password security and create very insecure passwords if they don't get any feedback. They recommend that users should be given training and instruction on how to construct usable and secure passwords. Advices for password selection such as *"Good passwords appear to be random characters. The wider the variety of characters the better. Mixing letters with numbers is better than letters alone. Mixing special characters with numbers and letters is better still."*, can be found on many websites [59]. These advices may be helpful in general, but users could just ignore them. To address these problems, strategies are needed which force or effectively recommend users to select stronger passwords.

3.1 Password Composition Policies

One approach to guide or force users to select stronger passwords is the use of password composition policies. These policies lay out the exact rules on how to create a strong password. Their aim is to add additional complexity into the password so that it gets a higher entropy. Such rules could be for example, that the password must have a minimum length of 8 characters, must contain at least one uppercase letter, digit and symbol, has to be changed every six months or is not allowed to contain a dictionary word. Theoretically an authentication system could require any rule which would be useful for its security. For example Shay et al. [47] used eight different policies in their study, among them *comp8*, which required the password to be at least 8 characters long and include a lowercase, uppercase, digit and symbol character, while the policy *basic16* required the password only to be at least 16 characters long.

The impact of password composition policies on password selection has been researched by several studies. Mazurek et al. [29] found that the use of password policies is strongly correlated with password strength. By using policies the selection of uppercase letters, digits and symbols strongly increased, which finally led to the creation of stronger passwords. Especially the addition of symbols and uppercase letters had the strongest effect on password strength, unless they were not placed at easy predictable places like uppercase letters as the first character and symbols at the end of the password. They also report that passwords of users who were annoyed with complying with the policy, were 46% more likely be guessed than those who did not report annoyance. This could be an indication that annoyance of password policies is counterproductive. In a study of Shay et al. [47] 28% of passwords which were created under the *comp8* policy fulfilled the symbol requirement by only adding a "!" at the end of the password and using no other symbols, while 54.8% of the passwords used an uppercase letter as the first character and using no other uppercase letter. Other studies also revealed that users often comply with policy rules by placing characters in predictable ways [57, 61]. Further studies, which measured the creation time, memorability and crackability of passwords under different policies, found that stricter policies can make passwords harder to crack but also harder to create and remember [26, 37, 48]. Such policies led users also to adapt circumvention strategies, for example requiring the use of at least 3 digits resulted in simply appending ”123” on the end of the password [21, 24, 45].

The possibility to use different rules raises the question of which one of them results in more secure passwords and at the same time provides memorable ones. Komanduri et al. [26] compared the two different password composition policies *comp8* and *basic16*. They found that the *basic16* policy resulted in significantly less predictable passwords than *comp8* and passwords were easier to remember and less likely to be written down. This was also observed by Kelley et al. [24]. Proctor et al. [37] also found that increasing the minimum length was more effective than applying content constraints. In a recent study Shay et al. [47] compared 8 different password policies to identify a policy which provides better usability and security benefits than the widely used *comp8* policy. In general they found that policies which require more length lead to more usability and sometimes more security than those who require only a mix of character types. They identified the two policies *3class12* (minimum 12 characters long and at least 3 types of different characters) and *2word16* (minimum 16 characters long and including 2 letter sequences which are seperated by a non-letter sequence) to be both more usable and stronger than *comp8*, while *3class12* provided more usability than *2word16* and *2word16* provided more security than *3class12*.

3.2 Password Meters

Password checkers or meters recommend users to create stronger passwords by giving them proactive feedback on the strength of their chosen passwords. They aim to increase the awareness on password security and finally encourage and nudge users to choose stronger passwords on their own. For example in a study of Ur et al. [51] a user mentioned: *”It kept me from being lazy when creating my password. [I] probably would not have capitalized any letters if not for the meter.”*. Password meters can also be combined with composition policies, but they can label passwords as weak which would be considered as strong by policies. For example ”P@$$w0rd” would fulfill all composition policies of *comp8*, but with a leeted dictionary this password could easily be derived from ”password”. During password creation the strength of the password is determined in real time and a scoring of the strength is given as a visual and/or textual feedback. Often the strength score is calculated by taking into account the password length, character set, known patterns like ”qwerty” and a dictionary check (blacklist) [13]. Parameters which are related to the user, such as his real/user name or email address, can also be included. Ur et al. [51] discovered three different types of meter visualizations. Bar-like meters, which are represented as a progress-bar metaphor or as a segmented-bar metaphor, checkmark or x-systems and text. Most commonly they identified password meters which are represented as colored bars and scale password strength from weak to strong by changing color and size from a short red bar to a long green bar and showing the user additionally a word which qualifies the password strength, for example ”weak”, ”fair” or ”strong” *(see figure 1)*.

The impact of password meters has been studied by the following studies. De Carnavalet et al. [13] evaluated the real life usage of password meters from eleven popular web services like Google, Dropbox, Yahoo!, Microsoft and others. Password meters from the web services widely differed in labeling the same passwords, for example ”Password1” was rated very weak by Dropbox but very strong by Yahoo! or ”football#1” was rated very weak by Dropbox but perfect by Twitter. Such inconsistencies were found often. This case effects usability and security, because if strong passwords are marked as weak than users might get frustrated which reduces usability while weak passwords marked as strong decreases security. They further discovered that strength scales and labels also varied from website to website, for example PayPal had three possible variations for the password strength (poor-medium-good), while Twitter had six (too short-obvious-not secure-could be more secure-okay-perfect). Likewise with password policies, they found that too strict password checkers annoyed users and made them less motivated to satisfy the meters. In general they suggest that the design of an ideal password checker should consider the use of inherent patterns in user choices, dictionaries used in cracking tools, exposures of large password databases and user adaptation against password policies [13].

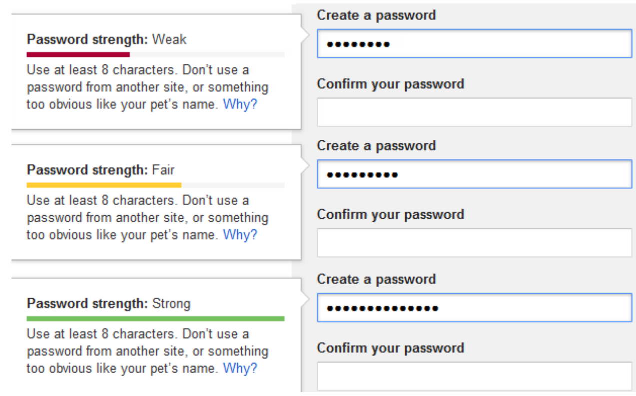


Fig. 1. Google Mail Password Meter tested 06/02/2014, ”Weak” = (”password”, ”helloworld1”), ”Fair” = (”P@$$w0rd1”, ”letmeinplease”), ”Strong” = (”XE%#6cQ{9*Ze”, ”1@maG3niu$”)

Ur et al. [51] tested the effect of 14 visually different password meters and found that meters influenced user behavior and led users to create significantly longer passwords, which were also significantly stronger and had no observable effect on their memorability. Passwords which were created without a meter were cracked at a higher rate than passwords created with the presence of meters. Users were more motivated to include symbols and additional lowercase letters to their passwords and they also changed their passwords more often while entering them. They were also more annoyed and unmotivated to satisfy stringent meters, while the more lenient meters led users more often choose passwords which were at least not labeled as weak. Some users complained that they did not understand why the password meter rated their password as weak and some said that meters encouraged and reminded them to use a more secure password. They also found that the visual appearance of the meters did not lead to significant differences of password strength or user behavior, but the combination of text and visualization had a better impact than either of them in isolation.

Egelman et al. [15] also found that password meters guide users towards selecting stronger passwords while not increasing memorability problems. But this also depends on the context of the passwords used. The passwords for unimportant accounts were not much influenced by the presence of a password meter while the passwords for important accounts were created significantly stronger. As well as Ur et al. [51] they observed that the mere presence of a meter has greater impact than individual meter design decisions. They further reported that if meters don't explain users why their selected passwords are considered as weak, they just try to make small modifications such as appending digits at the end to get a better score.

Castelluccia et al. [9] have made an approach to improve the effectiveness of password strength accuracy of password meters to avoid problems regarding false feedback about password strength. They suggest the use of Markov chains to develop an adaptive password strength-meter. The password strength is estimated by the probability of occurrences of the n-grams which are built and updated with each password contributed from every new user of a website. With this method the meter automatically adjusts to each website-specific tendencies and therefore prevents the clustering of passwords [9]. They could show that their approach outperformed existing schemes and it might be an approach to lead to more uniform meters in use.

3.3 Computer-generated Passwords

A method to ensure maximum theoretical password strength is to give users a fixed password which was randomly created by the system. This approach has the problem that these passwords are hardest to remember for the users and therefore not very usable [48, 59]. There are other approaches which bring passwords closer to randomly selected ones while providing usability.

Forget et al. [17] propose an approach towards more secure passwords by using Persuasive Text Passwords (PTP). The idea is that users can create their own passwords and PTP inserts 1 to 4 randomly generated characters to a random position. However, they observed that these randomly generated characters also caused memorability problems. The more additional random characters were inserted into the password, the simpler base passwords the users selected. They suggest that inserting two randomly generated characters would be the best tradeoff between security and memorability.

Houshmand et al. [20] developed a similar system, called analyze-modify password (AMP). AMP first analyzes a user chosen password whether it is weak or strong by estimating the probability of the password being cracked. If a password is weak, AMP will modify it by adding one random character at a random position, for example modifying "life45!" to "life^45!" [20]. After that the password will be once again analyzed and if it is still too weak then an additional random character will be inserted. With this method the original user chosen password should be preserved as much that it is still memorable for the user. In their experiments they could modify weak passwords, which first could be cracked with a success rate of 53% by the "John the Ripper" tool, to stronger passwords, which afterwards could only be cracked with a success rate of 0.27%.

Weir et al. [57] argue that the main problem with these approaches is that they only function effectively if they are applied to a limited number of passwords in use. Since these approaches insert characters randomly, passwords of different sites would all be different. This would be good for security, because users can not reuse the exact same password across multiple accounts. But the chances that users reuse the same base password and write their modified passwords down would be extremely high [57].

4 Novel Approaches

For now password composition policies and password meters may be the most used approaches to increase the practical password space. In this section novel approaches are presented. They are divided into three categories. First, feedback-based approaches which give users a novel way of advice or persuasive feedback during password creation. Second, input-based approaches which use or adapt the input interface. At last, graphical approaches which aim to replace the input of alphanumerical text with image based inputs. Since the focus of this paper is the use of alphanumerical passwords, the graphical approaches are just briefly introduced.

4.1 Feedback-Based

A strategy to create stronger and easily memorable passwords is the use of passphrases or mnemonic phrases [59]. An authentication system could recommend users to use these techniques. Passphrases are sequences of words which form a long password, for example *"WhenyouplaytheGameofThronesyouwinoryoudie"*. This would be a bad example for a strong password, because this is a quote from a popular series and an attacker could add such popular quotes to his dictionary. To address the use of popular phrases approaches have been made, where the system automatically assigns passphrases to the users, for example "correcthorsebatterystaple" [46]. Shay et al. [46] compared such passphrases with randomly generated passwords with similar entropy. They found some major usability problems. Passphrases and random passwords were forgotten at similar rates, they led to similar levels of user difficulty and annoyance, and were written down by most of the users. Users took significantly longer to enter them and they often required error correction to address entry mistakes. Similar results were found by Keith et al. [23]. The mnemonic phrase method advices the user to think about a meaningful and easy to memorize sentence and take the first character of every word from the sentence to build a password. For example a typical advice could be: *"A good technique for choosing a password is to use the first letters of a phrase. However, don't pick a well known phrase like - An apple a day keeps the doctor away - (Aaadktda). Instead, pick something like - My dog's first name is Rex - (MdfniR) or - My sister Peg is 24 years old - (MsPi24yo) "* [59]. Yan et al. [59] found that passwords based on mnemonic phrases were as easy to remember as naively selected passwords and at the same time as strong as randomly generated passwords. This would be a desirable approach to have both, strong security and good memorability.

Another feedback-based approach is a study of Schechter et al. [41], who propose to build an oracle for already existing passwords based on popularity. Users who select a password which has already been used a number of times by other users, get feedback that the use of this password is not allowed anymore. A problem is the use of the oracle without revealing the actual password to an attacker who could simply try any password.

A method developed by Vance et al. [53] uses "interactive fear appeals" to motivate users to increase the strength of their passwords. The idea of this novel approach is that users get real time messages, which convey the seriousness of a threat. The message informs the users how quickly their password can be guessed by a hacker, such as *"The password you entered is very insecure and may take a hacker 1 day to guess."* [53]. Additional advice of how to create a stronger password is also given *(see figure 2)*. They examined a study to compare their approach with interactive password strength meters as presented in section 3.2 and with a static form of fear appeal treatment. The results of their study show that the interactive fear appeal treatment resulted in significantly stronger passwords, while in contrast both other methods did not increase the password strength significantly. This finding is interesting regarding the interactivity of the approach, as the same fear appeal treatment in a static form did not lead to stronger passwords, whereas the interactive one did. Further interesting is that the interactive password strength meters did not increase password strength significantly, which is in contrast to findings mentioned in section 3.2. Vance et al. [53] suggest that the interactivity of password meters should be coupled with information about the severity and susceptibility of threats and the efficacy of the user's response, so that users can better understand the implications of their actions.

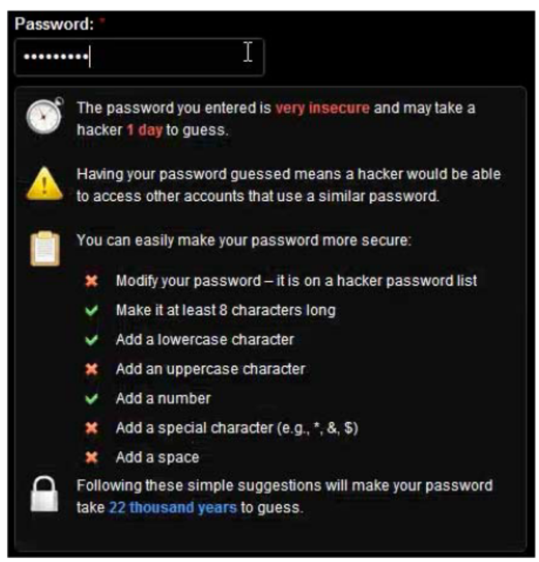


Fig. 2. Interactive fear appeal treatment [53]. Throughout password creation, users are shown a message which informs them how quickly their password can be guessed by a hacker. Users additionally get advice of how to create a more secure password.

Egelman et al. [15] propose a novel approach, which they call *"peer pressure motivator"*. It is based on social pressure by presenting password strength relative to all of the other users on the system. In their study they could show that the feedback significantly led to passwords of stronger entropy, even stronger than traditional password meters. Castellucia et al. [9] also imagined such an approach. They thought that the password strength score, which they calculated for their adaptive password strength meters, could also be shown by comparing users with their peers. *"Users could be prompted with a message that says, for example - your password is amongst the 5% weakest passwords on our website"* [9]. This might motivate users to create a stronger password. It could also be an incentive for users to reach the top 5%.

4.2 Input-Based

Approaches mentioned previously, focused on either enforcing or recommending users to create stronger passwords. These approaches are independent from the input interface. It would be interesting to explore whether the input interface or text entry method impacts password security and usability. A text entry method can be of a physical form, such as a hardware keyboard, buttons or sensors, or of a virtual/software-based form, such as a virtual keyboard of smartphones or tablets [28]. The design of text entry methods can strongly influence usability aspects, for example on different keyboard layouts it can significantly vary how quickly and effortlessly a character or phrase can be typed [28].

Yang et al. [60] examined whether the design of text entry methods affects the security of generated passwords. They supposed that users may generate passwords by using the characters on the display as generation cues. In particular the difficulty to reach a character could affect the probability of its inclusion in a password. For example on a laptop keyboard, numbers or symbols can be selected very quickly, while on a typical virtual smartphone display, users have to switch to a new view and often search for the desired symbols. This effort might affect the creation of passwords. To study these assumptions Yang et al. [60] tested different text entry methods on different groups. They found that basic structures and distributions of passwords were significantly different across groups. The more difficult it was to reach keys the more users chose characters from a smaller subset of characters. For example the group with the most difficult text entry method had much more passwords which contained only lowercase letters. As mentioned in section 2, a smaller subset of characters would result in less secure passwords. This fact has also been observed by them, as passwords across groups were significantly different against resistance to password cracking attacks. An attacker could make use of such vulnerabilities. For example he could adapt password cracking methods specifically for a website,which is accessed mostly in mobile context, to take advantage of the typical user choices on the mobile text entry method. Since they found that some text entry methods resulted in significantly less secure passwords, it would be reasonable that the design of text entry methods could also be adapted in a way to lead to the construction of more secure passwords. This approach would be an interesting way to possibly increase the practical password space. There has to be further research to explore this attempt.

Another approach of input-based strengthening of passwords is the PsychoPass method, which was proposed by Cipresso et al. [10] and further improved by Brumen et al. [7]. The idea is that, instead of remembering a specific string of characters as password, the user has to create, memorize and recall an action sequence on the keyboard. The applied steps are: *"(1) begin with a letter on the keyboard (2) memorize a sequence of actions (something like "the key on the left, then the upper one, then the one on the right", and so on) (3) memorize the sequence (not the letters used) (4) create as many passwords as you want by remembering only the first letter and the sequence"* [10]. Brumen et al. [7] analyzed the first approach by Cipresso et al. [10] and argued that there are two major problems. First, the method can only be produced and reproduced with the same keyboard layout. Second, the passwords would still be weak, because the passwords are based on proximity of the keys and therefore more predictable. They developed three improvements to make the PsychoPass method more secure. One improvement is to use she SHIFT and ALT-GR keys in combination with other keys. The second one is to use keys, which are more distances apart than only 1. The last improvement is that the number of keys in sequence should be at least 9 to 10. In a further study they could show that with these improvements passwords generated with the PsychoPass method are as resilient to brute force attacks than randomly generated passwords while being much easier to remember [6]. A 18 character long password would be comparable with a 11 character long, randomly generated password (*comp8*) [6].

One disadvantage of mobile devices is that the text entry can often be more time consuming and error prone than on traditional input devices such as hardware keyboards. Jakobsson et al. [22] introduced an approach called *fastwords*, which takes advantage of error correction and completion methods often used on mobile devices. They propose to create passwords as a sequence of dictionary words, because auto correction and completion methods can address dictionary words much better than usually passwords. However, the sum of the used dictionary words has to achieve a certain threshold of entropy to be accepted as a secure password. For example Jakobsson et al. argue that the *fastword* "frog word flat" would be a secure password, while "I love you honey" would not be accepted as a secure password. In their studies they could show that *fastwords* are about twice as fast to enter on mobile keyboards and three times as fast on full-size keyboards, while providing higher entropy and recall rates than usually chosen passwords.

4.3 Graphical-Based

Many approaches have been made in the field of graphical password authentication, for example *Draw-a-secret* or *PassFaces*, to make passwords more memorable and secure [3]. Users benefit from the psychological fact that pictures are better to be remembered than text [33]. Studies have shown that graphical passwords can actually provide a better security than alphanumerical passwords [3, 50]. Nonetheless, they also bring new challenges for the authentication system, for example the password registration and login process takes much longer [40]. Graphical-based passwords are more and more found in mobile context, but in particular here they are affected from further security problems, such as by smudge attacks or shoulder surfing [3, 56]. Similar to alphanumerical passwords, the security of graphical passwords is also affected by human factors which decreases the theoretical password space. For example attractive images of faces from the *PassFaces* approach were more often selected than others [3]. By analyzing a set of 506 user generated Android Patterns Saltuk et al. [39] found that only 2% of the users used the possibility to draw "knight-moves", which strongly decreases the theoretical password space.

5 Discussion & Conclusion

If people register on an online service, such as Gmail or Dropbox, it will come to two important decisions: *"Please select your username!: ..."* and most important *"Please select your password!: ..."*. The security and privacy of users' online data depends on the strength of the selected password, as the username is mostly public. Most users are not experts on password security. However, usually they are good at selecting practical passwords which are easy to remember, but unfortunately also easy to guess. To address both usability and security, methods have been presented which should guide users to select stronger passwords *(see table 1)*.

Password composition policies lead to the construction of stronger passwords, because they force users to use more complex passwords with mixed characters and a minimum length. Their disadvantage is that users often use simple strategies to comply with the rules (complying the uppercase letter requirement with the first character, the number requirement with appending them at the end, the symbol requirement with leet transformations or mostly using the same symbol). Although it has been shown that password policies lead to significantly stronger passwords, the user adaptations also decrease the password space, because an attacker can specifically adapt his cracking methods to the typical user behaviors. However, in the choice of which policy

Table 1. Advantages (⊕) and disadvantages (⊖) of the presented methods to increase the practical password space *(pw = password/s)*

Method	**Security**	**Usability**
Password Composition Policies [21, 24, 26, 29, 37, 45, 47, 48, 57, 61]	⊕ stronger pw ⊖ user adaptation, placement of characters on predictable places	⊕ easy to remember with lenient policies, harder to remember with stricter policies ⊖ no interactive feedback
Password Meters [9, 13, 15, 51]	⊕ proactive checking, less weak pw ⊖ depends on appropriate strength measuring	⊕ easy to remember ⊖ no pw improvement proposal, no uniform meters in use
Computer-generated (random) [48, 59]	⊕ maximum strong pw	⊖ hardest to remember and type
Analyze-Modify Password [17, 20]	⊕ extra pw strength	⊖ multiple pw hard to remember
Passphrases (automatically assigned) [23, 46] Mnemonic Phrases [59]	⊕ stronger pw (mnemonic comparable with random pw) ⊖ less secure pw due to popular phrases	⊕ easy to remember (mnemonic) ⊖ harder to remember(passphrases)
Persuasive Feedback [15, 41, 53]	⊕ stronger pw, raises security awareness ⊖ need for appropriate strength measuring	⊕ pw improvement proposal (Interactive fear appeal treatment), desirable incentive
Input-based [6, 7, 10, 22, 60]	⊕ stronger pw (PsychoPass pw appear as randomly generated pw) ⊖ less secure pw on reduced mobile keyboards	⊕ easy to remember, multiple pw with the same action sequence and different start keys ⊖ keyboard dependent

is the better approach there is again a tradeoff between usability and security. Policies with too strict requirements result in more secure passwords but in less memorable ones. More lenient policies lead to more memorable and less secure ones. But there are policies which achieve a good and sufficient balance between usability and security. Requiring longer passwords with less mixed characters seems to be the more promising way.

In contrast the advantage of password meters is that they can check passwords proactively. In this way they can address user adaptations, typical keyboard patterns like ”qwerty” and other easy to guess user behavior. They further benefit from included blacklists, which can easily be updated on a regular basis. The given feedback does indeed guide users towards selecting stronger passwords, because users tend to avoid passwords which are labeled as weak. Likewise with password policies, password meters which score password strength too strictly, lead to more unmotivated behavior and less effort to satisfy the meter. This is because users don’t understand why the meters label their passwords as weak. In my opinion the major disadvantage is that password meters don’t give appropriate feedback on how to construct or change passwords to achieve a better score. They should also explain why passwords have been rejected or labeled badly, otherwise users simply try to add additional numbers or symbols at the end of the password to reach a better score. One further drawback is that password meters in use widely differ in labeling same passwords. There is no uniform approach to measure password strength leading to user confusion and distrust. Password meters face the challenge to give a truthful estimation of password strength, which is unfortunately not easy determinable with user chosen passwords. A good approach would be to determine the strength score by using modern, up to date, cracking algorithms.

Computer generated approaches have the advantage that they always add some randomness into the passwords which guarantees more security. But their biggest disadvantage is that they are very hard to remember and therefore often written down. Especially it is a too high cognitive burden to force users to remember multiple random or partly random passwords.

Password creation strategies such as passphrases or mnemonic phrases can result in stronger passwords, if they are applied correctly. Mnemonic phrases can ensure passwords which are as secure as randomly generated passwords and at the same time easy to remember. This is a desirable approach. An authentication system can advice users to use these techniques. The challenge here is that there is no guarantee that users apply the recommendations correctly and therefore no guarantee for stronger passwords. In my opinion it would be a promising approach if such techniques are combined with interactive meters, which additionally use a dictionary check for popular phrases.

The persuasive feedback based approaches have the advantage that they raise the awareness for stronger passwords by giving a reasonable threat or an desirable incentive. This leads to more user motivation and finally to stronger passwords. The interactive fear appeal method has the further advantage that at the same time it gives users recommendations how to construct a better password. Their disadvantage is that they need appropriate strength measuring, otherwise they could lead to user frustration/annoyance if labeling strong passwords as weak or weak ones as strong.

Input based approaches have the advantage that users can remember action sequences instead of single passwords. Users can remember multiple different passwords with only remembering one action sequence and for each password another start key. The passwords additionally appear to be randomly generated because they contain no common dictionary words and therefore are much stronger. However, this could be a disadvantage if new dictionaries with popular action sequences emerge. Their major drawback is that the use of the passwords is dependent from the input keyboard and can’t be easily transfered to another keyboard.

An advantage of all presented approaches is that in one form or another they raise the awareness for security importance by simply being present. User education/instruction and interactive feedback and training seems to be a successful strategy and does indeed guide users towards selecting stronger passwords.

References

[1] A. Adams and M. A. Sasse. Users are not the enemy. *Commun. ACM*, 42(12):40–46, Dec. 1999.

[2] A. Adams, M. A. Sasse, and P. Lunt. Making passwords secure and usable. In *People and Computers XII*, pages 1–19. Springer, 1997.

[3] R. Biddle, S. Chiasson, and P. Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Comput. Surv.*, 44(4):19:1–19:41, Sept. 2012.

[4] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 538–552, May 2012.

[5] A. S. Brown, E. Bracken, S. Zoccoli, and K. Douglas. Generating and remembering passwords. *Appl. Cognitive Psych.*, 18(6):641–651, 2004.

[6] B. Brumen and A. Cernezel. Brute force analysis of psychopass-generated passwords. 2014.

[7] B. Brumen, M. Heričko, I. Rozman, and M. Hölbl. Security analysis and improvements to the psychopass method. *Journal of medical Internet research*, 15(8), 2013.

[8] W. E. Burr, D. F. Dodson, and W. T. Polk. *Electronic authentication guideline*. Citeseer, 2004.

[9] C. Castelluccia, M. Dürmuth, and D. Perito. Adaptive password-strength meters from markov models. In *Proc. NDSS*, volume 2012, 2012.

[10] P. Cipresso, A. Gaggioli, S. Serino, S. Cipresso, and G. Riva. How to create memorizable and strong passwords. *J Med Internet Res*, 14(1):e10, Jan 2012.

[11] L. F. Cranor and S. Garfinkel. Guest editors’ introduction: Secure or usable? *Security & Privacy, IEEE*, 2(5):16–18, 2004.

[12] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The tangled web of password reuse. 2014.

[13] X. de Carné de Carnavalet and M. Mannan. From very weak to very strong: Analyzing password-strength meters. In *Network and Distributed System Security (NDSS) Symp. 2014*. Internet Society, February 2014.

[14] M. Dell'Amico, P. Michiardi, and Y. Roudier. Password strength: An empirical analysis. In *INFOCOM, 2010 Proc. IEEE*, p. 1–9, March 2010.

[15] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. Does my password go up to eleven?: The impact of password meters on password selection. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2379–2388, New York, NY, USA, 2013. ACM.

[16] D. Florencio and C. Herley. A large-scale study of web password habits. In *Proc. of the 16th International Conference on World Wide Web*, WWW '07, pages 657–666, New York, NY, USA, 2007. ACM.

[17] A. Forget, S. Chiasson, P. C. Oorschot, and R. Biddle. Persuasion for stronger passwords: Motivation and pilot study. In *Proc. of the 3rd International Conference on Persuasive Technology*, PERSUASIVE '08, pages 140–150, Berlin, Heidelberg, 2008. Springer-Verlag.

[18] R. Grenoble. 'secret' nuclear missile launch code during cold war was '00000000'. Retrieved from http://www.huffingtonpost.com/2013/12/05/nuclear-missile-code-00000000-cold-war_n_4386784.html, Jan 2014, online accessed 06/01/2014.

[19] S. Haque, M. Wright, and S. Scielzo. A study of user password strategy for multiple accounts. In *Proc. of the third ACM conference on Data and application security and privacy*, pages 173–176. ACM, 2013.

[20] S. Houshmand and S. Aggarwal. Building better passwords using probabilistic techniques. In *Proc. of the 28th Annual Comp. Sec. Applications Conf.*, ACSAC '12, pages 109–118, New York, NY, USA, 2012. ACM.

[21] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: Password use in the wild. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 383–392, New York, NY, USA, 2010. ACM.

[22] M. Jakobsson and R. Akavipat. Rethinking passwords to adapt to constrained keyboards. *Proc. IEEE MoST*, 2012.

[23] M. Keith, B. Shao, and P. J. Steinbart. The usability of passphrases for authentication: An empirical field study. *International journal of human-computer studies*, 65(1):17–28, 2007.

[24] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proc. of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 523–537, Washington, DC, USA, 2012. IEEE Computer Society.

[25] D. V. Klein. Foiling the cracker: A survey of, and improvements to, password security. In *Proc. of the 2nd USENIX Security Workshop*, pages 5–14, 1990.

[26] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: Measuring the effect of password-composition policies. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2595–2604, New York, NY, USA, 2011. ACM.

[27] W. Ma, J. Campbell, D. Tran, and D. Kleeman. A conceptual framework for assessing password quality. *IJCSNS International Journal of Computer Science and Network Security*, 7(1):179–185, 2007.

[28] I. S. MacKenzie and K. Tanaka-Ishii. *Text entry systems: Mobility, accessibility, universality*. Morgan Kaufmann, 2010.

[29] M. L. Mazurek, S. Komanduri, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur. Measuring password guessability for an entire university. In *Proc. of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 173–186, New York, NY, USA, 2013. ACM.

[30] A. Narayanan and V. Shmatikov. Fast dictionary attacks on passwords using time-space tradeoff. In *Proc. of the 12th ACM conference on Computer and communications security*, pages 364–372. ACM, 2005.

[31] Openwall. John the ripper password cracker. Retrieved from http://www.openwall.com/john/ , online accessed 06/01/2014.

[32] Openwall. Openwall wordlists collection. Retrieved from http://www.openwall.com/wordlists/ , online accessed 06/01/2014.

[33] A. Paivio, T. Rogers, and P. C. Smythe. Why are pictures easier to recall than words? *Psychonomic Science*, 11(4):137–138, 1968.

[34] N. Perlroth. Ebay urges new passwords after breach. Retrieved from http://www.nytimes.com/2014/05/22/technology/ebay-reports-attack-on-its-computer-network.html, May 2014, online accessed 06/01/2014.

[35] D. Pescovitz. Top 500 worst passwords. Retrieved from http://boingboing.net/2009/01/02/top-500-worst-passwo.html, Jan 2009, online accessed 06/01/2014.

[36] C. P. Pfleeger and S. L. Pfleeger. *Security in computing*. Prentice Hall Professional, 2003.

[37] R. W. Proctor, M.-C. Lien, K.-P. L. Vu, E. E. Schultz, and G. Salvendy. Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods, Instruments, & Computers*, 34(2):163–169, 2002.

[38] S. Riley. Password security: What users know and what they actually do. *Usability News*, 8(1):2833–2836, 2006.

[39] O. Saltuk and I. Kosan. A human model for guessing android patterns. *Ludwig Maximilians University of Munich*, June 2014.

[40] F. Schaub, M. Walch, B. Könings, and M. Weber. Exploring the design space of graphical passwords on smartphones. In *Proc. of the Ninth Symposium on Usable Privacy and Security*, page 11. ACM, 2013.

[41] S. Schechter, C. Herley, and M. Mitzenmacher. Popularity is everything: A new approach to protecting passwords from statistical-guessing attacks. In *Proc. of the 5th USENIX conference on Hot topics in security*, pages 1–8. USENIX Association, 2010.

[42] ScrapMaker.com. Rockyou. Retrieved from http://scrapmaker.com/view/dictionaries/rockyou.txt , online accessed 06/01/2014.

[43] C. E. Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

[44] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[45] R. Shay, A. Bhargav-Spantzel, and E. Bertino. Password policy simulation and analysis. In *Proc. of the 2007 ACM Workshop on Digital Identity Management*, DIM '07, pages 1–10, New York, NY, USA, 2007. ACM.

[46] R. Shay, P. G. Kelley, S. Komanduri, M. L. Mazurek, B. Ur, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proc. of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 7:1–7:20, New York, NY, USA, 2012. ACM.

[47] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, and L. F. Cranor. Can long passwords be secure and usable? In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2927–2936, New York, NY, USA, 2014. ACM.

[48] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: User attitudes and behaviors. In *Proc. of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 2:1–2:20, New York, NY, USA, 2010. ACM.

[49] E. H. Spafford. Observing reusable password choices. *Proc. of the 3rd USENIX Security Workshop*, 1992.

[50] X. Suo, Y. Zhu, and G. S. Owen. Graphical passwords: A survey. In *Proc. of the 21st Annual Computer Security Applications Conf.*, ACSAC '05, pages 463–472, Washington, DC, USA, 2005. IEEE Computer Society.

[51] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How does your password measure up? the effect of strength meters on password creation. In *Proc. of the 21st USENIX Conference on Security Symposium*, Security'12, pages 5–5, Berkeley, CA, USA, 2012. USENIX Association.

[52] B. Ur, S. Komanduri, R. Shay, S. Matsumoto, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, M. L. Mazurek, and T. Vidas. Poster: The art of password creation. In *Proc. of the IEEE Symp. on Security and Privacy*, volume 6, 2013.

[53] A. Vance, D. Eargle, K. Ouimet, and D. Straub. Enhancing password security through interactive fear appeals: A web-based field experiment. In *Proc. of the 2013 46th Hawaii International Conference on System Sciences*, HICSS '13, pages 2988–2997, Washington, DC, USA, 2013. IEEE Computer Society.

[54] R. Veras, C. Collins, and J. Thorpe. On the semantic patterns of passwords and their security impact. In *Network and Distributed System Security Symposium (NDSS14)*, 2014.

[55] E. Von Zezschwitz, A. De Luca, and H. Hussmann. Survival of the shortest: A retrospective analysis of influencing factors on password composition. In *Human-Computer Interaction–INTERACT 2013*, pages 460–467. Springer, 2013.

[56] E. von Zezschwitz, A. Koslow, A. De Luca, and H. Hussmann. Making

graphic-based authentication secure against smudge attacks. In *Proc. of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 277–286, New York, NY, USA, 2013. ACM.

[57] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proc. of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pages 162–175, New York, NY, USA, 2010. ACM.

[58] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 391–405. IEEE, 2009.

[59] J. J. Yan, A. F. Blackwell, R. J. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Security & privacy*, 2(5):25–31, 2004.

[60] Y. Yang, J. Lindqvist, and A. Oulasvirta. Text entry method affects password security. *CoRR*, abs/1403.1910, 2014.

[61] M. Zviran and W. J. Haga. Password security: an empirical study. *Journal of Management Information Systems*, 15:161–186, 1999.

Memorability in the design and the evaluation of knowledge-based authentication systems

Malin Eiband

Abstract— Alphanumeric passwords are still the most commonly used knowledge-based authentication mechanism although they have significant usability shortcomings of which memorability is the most serious one. Usable security research has therefore brought forth a variety of improvements and alternatives to traditional password authentication.
In this paper, we address the topics of human memory and memorability in the context of the design and the evaluation of these alternative authentication systems. We focus on the systems that have most often been discussed in recent scientific literature, including graphical authentication systems, passphrases and cognitive passwords. Since understanding the psychological background is crucial in order to design and evaluate memorable authentication systems, we first give a detailed overview of information processing in human memory. We present special aspects of memory that we consider interesting for usable security research, such as our ability to retrieve pictures and meaningful sentences more easily than random-looking strings. We then suggest how the psychological findings could be applied to knowledge-based authentication systems. Based on these insights, we analyze how the systems mentioned above take into account the different aspects of memory in their design and evaluation. We find that these aspects and memorability in general play indeed an important role in all of the considered systems, but that there are currently no best practices when it comes to designing user studies for the evaluation of memorability and that user studies are always conducted without the help of psychologists. We conclude our explanations with a discussion of current evaluation practices and give suggestions for future approaches.

Index Terms—usable security, memory, evaluation, password design, authentication systems

1 Introduction

Human memory is a most amazing thing: It lets us learn how to ride a bicycle, to speak foreign languages and to dream, permits us to recall experiences we made half a lifetime ago and to solve complex problems. It allows us to reason, to make decisions and to be creative. Yet it fails us when it comes to remembering the person's name we have just shaken hands with, where we parked our car – or what our email password was.

Alphanumeric passwords belong to the class of authentication systems that are based on the knowledge of the users: They require them to retrieve a particular piece of information from memory during the authentication procedure. Usable security research has made clear that even though traditional passwords are still one of the most common knowledge-based authentication systems, they have usability shortcomings which arise from human memory limitations [5, 14, 33]: A long, random-looking string for password authentication is more secure, but difficult to remember. Users therefore tend to choose short, meaningful passwords which are easy to remember but also easy to guess. Coping with multiple passwords leads to password reuse across several accounts as it increases the cognitive load and the passwords are likely to get confused in memory [31]. The ease of remembering a particular item of information thus directly affects the security of a knowledge-based authentication system. Therefore, the memorability of these systems has become an important issue in usable security research.

What do we actually mean when we talk of memory and and memorability? Memory is a concept taken from cognitive psychology, the research area that is concerned with how we perceive, store and remember information, and refers to a set of mechanisms we use to process data items [7]. Memorability, on the other hand, is a usability criterion introduced by Nielsen in 1993 [40] representing the ease of remembering a system when the user returns to it after a certain period of time. In terms of knowledge-based authentication systems, memorability describes how easily the users are able to retrieve the piece of information from memory that is required for successful login, at regular intervals over a long period of time, for instance even after a longer holiday [62].
A knowledge-based authentication system with good memorability should therefore be built in a way that it facilitates the access of the required item of information in memory during the authentication procedure, that is, the *design* of the system itself should encourage memorability. In this context, we can see two main approaches in current usable security research:

- Systems and authentication mechanisms that are designed to improve current practice by increasing the memorability of alphanumeric passwords, such as passphrases.
- Systems with an alternative design to alphanumeric passwords, such as graphical authentication systems.

Apart from the design, *evaluating* the actual memorability of a system in user studies is another important issue. Although memorability is sometimes deduced from the measured input speed, the error rate and the number of password[1] resets [21, 40, 60], it is dependent on many other factors such as the context in which a particular piece of information is retrieved from memory, the confusion of several items of information, or the forgetting of information. It is notable that the design of the user study itself can influence memorability and long-term memory effects [57].

In this paper, we first give a detailed overview of the prevailing modern memory theories in Section 2. We explain how encoding, storing and retrieving of information works and present techniques to transfer information to long-term memory as well as several aspects of human memory that we consider interesting for usable security research. In Section 3, we suggest how these findings could be applied to a system's design and the design of user studies. We then take a closer look at current authentication systems and analyze whether in their design and evaluation they adequately take into account the different aspects of memory and classify them with regard to the aspects of

- *Malin Eiband is studying Media Informatics at the University of Munich, Germany, E-mail: malin.eiband@campus.lmu.de*

- *This research paper was written for the Media Informatics Advanced Seminar 'Special Aspects of Usability', 2014*

[1]For the purpose of readability, we will use the term "password" for knowledge-based authentication systems in general while we will explicitly refer to traditional passwords as "alphanumeric passwords".

memory they are using. In the last section, we discuss current design and evaluation practices for knowledge-based authentication systems and give suggestions for future approaches.

2 Human memory and memorability

In order to consider memorability in the design and the evaluation of knowledge-based authentication systems, it is crucial to understand the basic memory processes as well as the particular aspects of human memory. In the following sections, we therefore take a closer look at the results cognitive psychology has brought forth over the past 60 years.

2.1 Information processing and basic cognitive processes

Cognitive psychologists mainly define memory as a set of mechanisms consisting of the three core operations of encoding, storing and retrieving of information [7], each representing a different stage of information processing [51]. Although there are models which do not include the theory of storage, researchers mutually agree on the concepts of encoding and retrieving of information and have found them to be highly interdependent [61]: Failure of encoding of information will make recollection difficult and equally, properly encoded information will be remembered more easily. Below, we go through these stages and elaborate thereby on three of the most influential theories of memory: the traditional three-store Atkinson-Shiffrin model, Baddeley's model of the working memory and Craik and Lockhart's levels-of-processing approach. The structure of our explanations is based on [61].

2.1.1 Encoding

Encoding is the initial step of information processing where we transform information gathered through our senses into a mental representation. It provides the cues, or retrieval paths, which we can later use for recall [34].This step is crucial for the success of mnemonic techniques of which we will talk later in section 2.2.

A very influential theory that sees encoding as the key to successful recollection is the levels-of-processing approach by Craik and Lockhart [19]. They argue that deep, meaningful information processing leads to better recall than superficial, sensory processing. Information processing is not limited to three storage locations as proposed in the Atkinson-Shiffrin model, but can theoretically have infinite levels of depth. Deeper understanding of information such as analyzing its meaning, relating it to other things or putting it into a context induces more retrieval paths in memory, which are essential for later recall [18].

2.1.2 Storage

Storage describes the holding of the encoded information in memory. This concept derives from the so-called *information-processing approach* in the mid-fifties which compares our mental processes to the processing of information in a computer with several stages [34]. Many theories about memory suggest that there are several storage locations with different characteristics respectively. The most popular model presented at that time, which has dominated research for many years, is the traditional three-store model by Atkinson and Shiffrin [6]. It divides memory into three distinct storage locations: sensory memory, short-term memory and long-term memory. Note that Craik and Lockhart's levels-of-processing approach does not include the concept of storage as it does not divide memory into separate storage locations. In this approach, the depth of processing rather than a certain storage location determines the ease of recall.

Sensory memory represents the interface to our environment: Stimuli are gathered through our senses and held in sensory memory for a very short time, up to two seconds, before they are forgotten – it is therefore the most fragile of the three memory stages.

If we pay attention to certain stimuli, the information associated with these stimuli passes from sensory memory on to short-term memory. Its capacity is limited: Miller has shown that short-term memory can hold five till nine items [36].

Moreover, material is stored in short-term memory by its sound rather than by its visual appearance. Short-term memory holds the data we are currently using for about thirty seconds, unless it is somehow repeated.

Subsequent researchers however have questioned the clear distinction between short- and long-term memory and the role of short-term memory as a passive information store. Instead, it is now assumed that short-term memory is a part of long-term memory rather than a separate storage location and does not simply store data, but actively manipulates the information it holds. Introduced by Baddeley and Hitch in 1974 [8], the so-called working-memory model is the most prevailing model today, so that short-term memory is now typically called working memory [51].

A fraction of the information stored in short-term memory finally reaches long-term memory. Research has found evidence that most information in long-term memory is encoded either semantically [28], visually [11] or acoustically [39], but the capacity of long-term memory and how long data is stored there remain unknown until today [51]. Since long-term memory contains both childhood memories as well as very recent information, long-term capacity seems to be enormous, maybe unlimited, and relatively permanent [34].

Depending on the nature of stored information, Squire has subdivided long-term memory into *declarative* (or *explicit*) memory and *procedural* (or *implicit*) memory [48]: Our declarative memory holds the organized knowledge we have gathered throughout our lifetime, such as learned facts or information about events we have experienced. It can be declared, that is, we can consciously access the stored items. Procedural memory holds information about how we do something, for example how we ride a bicycle, how we tie our shoes and so on. It is non-declarative, and can thus only be accessed through performing a particular action.

2.1.3 Retrieval

In retrieval, we localize stored information and recollect it from memory. As already mentioned above, retrieving and encoding of information are highly interdependent processes: We use the cues, or retrieval paths, we created during encoding to access information. It is important to note that recall will be more successful for cues that distinctively identify information than for more general cues that point to a whole set of data [61]. Retrieval can be subdivided into three main categories that emphasize the significance of cues: free recall, cued recall and recognition. Free recall is the process of remembering a particular information while no or only a general cue is provided. In cued recall on the other hand, a specific cue supports the retrieval of information. In recognition, cues are even more assistant: We do not have to retrieve a particular information, but are requested to judge whether we can remember a given item or not [30]. Research has demonstrated that retrieving information through recognition is almost always more successful than through free recall [29].

2.2 Transfer of information from short-term memory to long-term memory

How do we move information from short-term to long-term memory? Atkinson and Shiffrin already suggested that we can influence this process through *control processes*, such as rehearsal. Rehearsal and repetition are indeed means by which we can regulate the information flow from short-term to long-term memory. Rehearsal can be elaborative or maintaining: In elaborative rehearsal, we somehow elaborate the information we want to remember, that is, we either make meaningful connections between newly learned items and thus increase their memorability or we integrate the items into our existing patterns and knowledge, a process which is called *consolidation*. In maintenance rehearsal, we simply repeat the items over and over again, but do not elaborate their meaning [52].

The success of these rehearsal techniques depends on whether declarative or procedural memory is involved: For declarative memory, information can be transferred to long-term memory through elaborative rehearsal. Maintenance rehearsal would only temporarily keep the information in short-term memory, but not store it in long-term memory. For procedural memory on the other hand, we can indeed use maintenance rehearsal to achieve long-term storage, for instance in order to practice a certain movement [12].

Another important aspect in the context of long-term memory is organization – stored memories are organized memories [51]. There are various strategies we may use to organize information we want to remember and thus to support rehearsal [58], for instance by means of *mnemonics*. Mnemonics are mental activities known to improve encoding and retrieval of lists of items [38]. These techniques allow us to add meaning to information that would seem to be arbitrary otherwise, and hence, to create a useful retrieval cue. Well known examples for mnemonic techniques are the *first-letter technique*, where the first letters of words in a list build an acronym which we can later access to retrieve all list items, or *mental imagery*, where we form a vivid mental image of the list items interacting with each other.

2.3 Special aspects of memory

There are various remarkable aspects to human memory toward which encoding, storing and retrieving of information are biased, and which have proven to be rather robust. In the following sections, we present those aspects which we consider the most interesting for usable security research, and explain them in more detail.

2.3.1 Memory effects

Feedback effect. In memorizing information, we want to store correct items of information rather than wrong ones. It is possible during retrieval to produce errors that might persist even though we later try to overwrite them with the correct piece of information [32]. Corrective feedback however seems to prevent the storing of errors. It is not crucial whether the feedback is delayed or given directly [42].

Spacing effect. Our memories tend to be more accurate and long-lasting if we space learning sessions at certain intervals rather than study in sessions that lie closely together [26]. These learning practices are called *distributed practice* and *massed practice* respectively. The effect of distributed practice is dependent on the delay between learning and retrieval [42]: The interval between the learning sessions should ideally be between ten and twenty percent of the retrieval delay. For instance, if a piece of information should be retrieved after a delay of ten days, the sessions should lie one or two days apart. In general, greater distribution of sessions over time tends to be superior to shorter delays between learning sessions in order to boost long-term effects [9, 16, 17, 45, 51].

Pictorial superiority effect. The pictorial superiority effect is one of the best-studied memory effects. It describes our distinctive ability to remember even vast amounts of pictures better than their verbal representations [22, 49, 50]. This effect holds true not only for recognition, but also for free recall [41].

Testing effect. Taking a test after having learned information leads to better long-term recall [44] and, combined with immediate feedback, has even been demonstrated to be more effective than an additional learning trial [15]. Also the way in which knowledge is tested influences information processing: Short-answer tests where we have to produce the answer from memory by ourselves have proven superior to multiple-choice tests [35].

Von Restorff effect. Our memory tends to favor the unusual and unexpected: Items of information that stand out from their environment are more likely to be remembered than others [23, 56]. For instance, in a string that consist entirely of characters except for one digit, the digit will be remembered best compared to the rest of the string.

Generation effect. If we generate our own cues for retrieval, they are much more effective than retrieval cues that were provided for us [47]. This effect has been demonstrated to be independent of the retrieval condition, that is, it holds true for free recall and cued recall as well as recognition.

Self-reference effect. We remember information better if we relate that information to ourselves and our own personal experience [24]. The self-reference effect may be seen as a special case of the generation effect that takes advantage of the "self" providing a vast amount of distinctive memory cues [34].

2.3.2 Other influences on memory

Memory interference and memory decay. When we transfer information from short-term to long-term memory, our memories are susceptible to interference when already stored items and new, similar information compete and retrieval paths are displaced. Moreover, stored information gradually disappears over time – we simply forget memories as time passes, a phenomenon called memory decay [51].

The seven sins of memory. Schacter [46] has presented seven specific ways in which memory distortions limit the accurateness of our memory: transience (the forgetting of information over time); absent-mindedness (the forgetting of information due to missing attention during encoding or retrieval); blocking (items are temporarily inaccessible even though they are "on the tip of our tongue"); misattribution (we misattribute the origin of a piece of information); suggestibility (we create false memories because of suggestion); bias (encoding and retrieval of information are influenced by previous experiences or pre-existing beliefs) and persistence (we remember information we would like to forget).

Encoding specificity. Memory is context-dependent. Research has demonstrated that memory accuracy is often improved when the contexts of retrieval and encoding are similar. Context consists of internal and external factors. Internal context includes features such as emotions at the time of encoding, moods or states of consciousness. External context contains for instance the location where a certain piece of information was encoded or the activity we were engaged in while encoding [27]. Hence, when we encode information in a certain context, for example in a particular mood or location, we may recall that information better in a similar retrieval context. This relationship between encoding and retrieval is called *encoding specificity* [54].

Everyday memory. Everyday memory is not like memory that is tested in the laboratory. For instance, while we accomplish a well-defined task in the laboratory that limits the amount of information we have to search, in everyday life, our memory has to filter useful knowledge out of an enormous amount of information, so performance is likely to differ. Another example is context-dependence: Context-dependent memory can be demonstrated more easily in real life than under laboratory conditions where context-effects are often inconsistent [7]. If the evaluation of memory and memorability takes place in a laboratory, it is therefore important to be aware of these differences [37].

The influence of task. Effectiveness of encoding methods depends on the nature of the memory task required for retrieval [43]. When choosing a method for encoding information, we should not only choose the strategy that seems best for storing items in long-term memory, but also consider the purpose for retrieving the information [53]. Memory failures during user studies are largely a result of retrieval rather than storage failures. The nature of the task also influences context-dependence which is more likely to occur during recall tasks.

The difficulty to control cognitive processes. Task results in a user study may be influenced by the study participants. As our cognitive processes are active rather than passive, people tend to transform task instructions rather than passively obey researchers. This is the case especially for very robust memory effects like the self-reference effect [34].

3 Memorability in Usable Security Research

In this section, we take a closer look at the importance of cognitive psychology for usable security research. We will first apply the scientific findings discussed in the previous section to the design and the evaluation of knowledge-based authentication systems. Then, we will analyze whether the findings of cognitive psychology are adequately taken into account in current systems and classify these systems with regard to the aspects of memory they are using.

3.1 Mapping psychology onto usable security research

How can the psychological knowledge from Section 2 actually be applied to usable security research? How can it be made relevant for

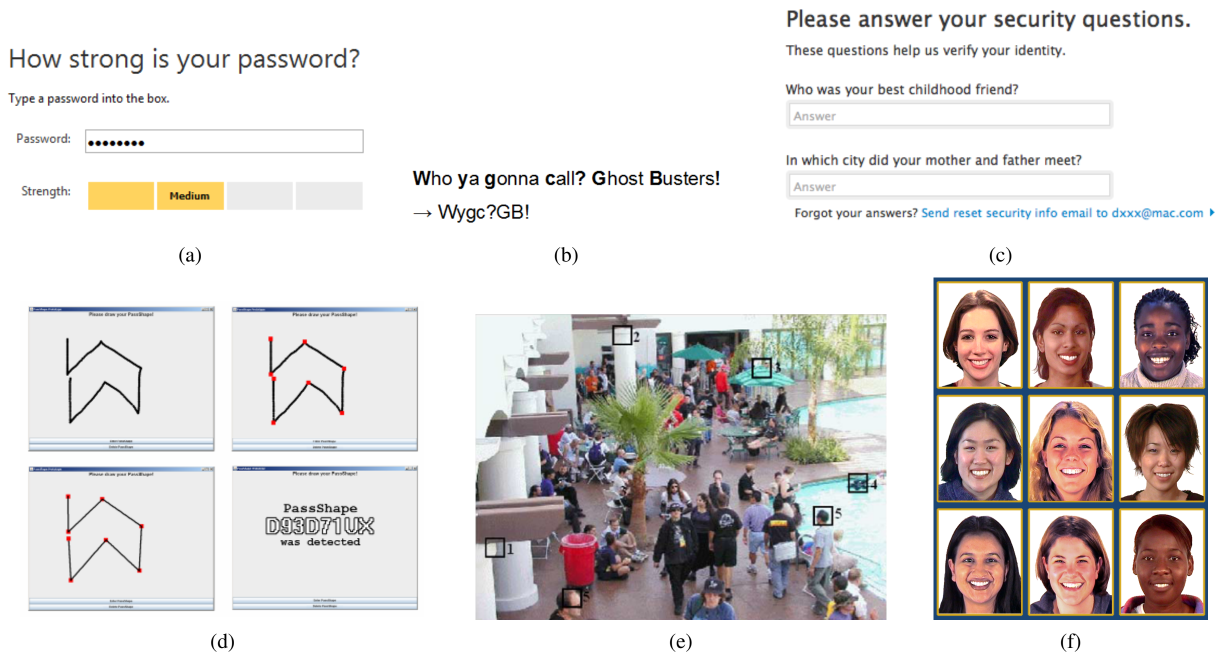


Fig. 1: (a) Microsoft's password checker [2]. In proactive password checking, a password is validated during password selection. (b) An example sentence we invented and its corresponding passphrase. The passphrase is created by taking the first letter of each word and the punctuation marks. (c) Apple's security questions are an example for cognitive passwords [1]. The correct answer to both questions grants access to the system. (d) The authentication process using passshapes [59]. In order to log in, the users have to draw a geometric shape that is then parsed into an alphanumeric string. (e) An example for a user password in the PassPoints system [60]. During authentication, the users have to click on the points they have selected as a password in the right order. (f) An example passfaces grid [3]. During authentication, several of these grids are displayed successively and the users have to identify their passfaces one at a time.

the context of knowledge-based authentication? We have seen that information processing is a complex mechanism but also that there are several aspects of memory that could be specifically useful for the design and the evaluation of memorable authentication systems. In the following sections, we suggest how these findings could be transferred to usable security research.

3.1.1 Designing memorable systems

Proper encoding of information and the organization of knowledge seem to be the key to successful retrieval in the case of declarative memory. A systems design should therefore encourage deep, meaningful processing of a newly learned password in order to support memorability, for example by means of mnemonic techniques. Simple repetition in contrast, that is, maintenance rehearsal, does not help to memorize new passwords. The *self-reference effect* could also be useful in this context, but one should bear in mind that this effect can also lead to passwords that are easily guessable and thus have a negative influence on password security. Moreover, according to the *generation effect*, letting the users select their password themselves instead of generating it for them should increase memorability.

During password selection, the *Von Restorff effect* could support password learning if the new password is somehow visually highlighted. Moreover, in order to prevent multiple passwords from interfering in memory, the systems design should encourage the creation of cues that distinctively identify the respective passwords. This could be especially interesting for systems that are based on cued recall where each cue could be specifically used to retrieve a particular password. Another important aspect of memory is the *pictorial superiority effect*. An authentication systems that relies on pictures rather than words should provide better memorability compared to a system that uses verbal information. This effect is the basis for graphical knowledge-based authentication systems that will be treated in section 3.2

On the other hand, there are several interesting aspects to procedural memory: A new password could be learned implicitly by means of pure repetition. Over time, after a number of inputs, a password might even be remembered better. Moreover, as procedural memory is non-declarative, an authentication system that takes advantage of procedural memory would keep users from possibly compromising security by writing down their passwords.

3.1.2 Designing user studies for the evaluation of memorability

An authentication system that considers memorability aims at creating long-term memory effects. The design of user studies that evaluate the actual memorability of the system should therefore also address long-term memory. As we have seen in section 2.3.1, the *spacing effect* is a suitable means by which long-term effects can be achieved if there are several learning sessions and if the intervals between sessions are sufficiently large. A long-term user study for the evaluation of memorability should thus always consist of various distributed sessions over time, lest one address only short-term memory. If the evaluated authentication system involves procedural memory rather than declarative memory, maintenance rehearsal is likely to produce long-term effects, otherwise, simple repetition of the new password during the user study is not enough. Instead, the user study should encourage the participants to generate their own retrieval cues, possibly by the help of mnemonics, and thus to process the new password more deeply. Furthermore, testing the new password after a short delay after password selection has been demonstrated to positively influence memorability [57]. Immediate repetition of the password during password selection, however, has no effect on memorability. After the test, feedback as to the correct answer should be given to prevent the participants from creating incorrect memories.

Moreover, the style of the tasks the participants have to perform influences the results of the study. Participants will be likely to retrieve information better during a recognition task rather than in free recall. Since everyday memory performance differs from the memory performance under laboratory conditions, we also suggest to undertake field studies in which the memorability of a system is tested under real-life conditions. For instance, the system could be included in the normal workflow of a participant in order to ensure a realistic test setting. It is important, however, that the password selection and the encoding of the new password take place under conditions similar to the conditions under which the password is to be retrieved at a later time, so that encoding specifity can be exploited.

Furthermore, possible interference in memory and memory decay as well as Schacters seven sins of memory [46] should always be kept in mind when conducting a user study on memorability and analyzing the results afterwards. It is also important to be aware of the participants, deliberately or involuntarily, transforming task instructions.

3.2 A classification of current knowledge-based authentication systems with respect to memorability

In this section, we present a number of examples of current research on the "password problem" [60], that is, the conflicting requirements of password security and password usability, and on the resulting need for more usable password systems. Our aim is not to compare the systems or to give advice as to when to use which system. We rather want to emphasize *if* the special aspects of memory we presented in Section 2.3 and memorability as a usability criterion have been adequately taken into account in the design and the evaluation of current knowledge-based authentication systems, and if so, *in which way*. In general, we see two main approaches to the usability shortcomings of traditional alphanumeric passwords which we will look at more closely:

- Systems and practices that try to improve the memorability of alphanumeric passwords. We will pay special attention to proactive password checking and pass phrases.

- Systems that represent alternatives to alphanumeric passwords. As graphical password systems have received significant attention, we will focus on these systems, including drawmetric (recall-based), locimetric (cued recall-based) and cognometric (recognition-based) graphical password systems [20]. As an example for a second alternative approach to traditional passwords, we will then present cognitive passwords.

Table 1 and 2 give an overview of the results of our analysis.

3.2.1 Proactive password checking

Design. Proactive password checking is not a password system itself, but a method that allows one to determine whether a password created by the user satisfies certain specified restrictions of a particular system [10]. An example for proactive password checking is shown in Figure 1a.

Analysis of the design. The design of this method takes advantage of the generation effect.

Evaluation. Vu et al. have conducted a one-week user study for proactive password checking that also evaluated memorability: Participants were instructed to create passwords for multiple accounts according to certain password restrictions in the first session and were told that they would have to remember them and that they would be granted a maximum of ten login attempts for each account. After a break of five minutes to distract short-term memory, participants then had to enter their passwords for all accounts. One week later, the retrieval procedure was repeated. The participants were all students and the user study took place under laboratory conditions [57].

Analysis of the evaluation. Vus user study is one of the few studies that involve multiple passwords and probable password interference. Moreover, the evaluation considers the spacing effect and the generation effect. The testing effect is exploited as well as users had to reenter their passwords after a short break. In subsequent studies, Vu has also combined proactive password checking with mnemonic methods like passphrases [57].

3.2.2 Passphrases

Design. Passphrases is the name of a method that asks the user to generate a password by formulating a sentence and then build the password by taking the first letter of each word, including use of capital and small letters. (see Fig. 1b).

Analysis of the design. Passphrases encourage deeper processing of information than traditional passwords through the use of the mnemonic "first- letter technique". They also provide a retrieval clue for subsequent recall, that is, users do not have to remember the random-looking password but rather the meaningful sentence from which they can reconstruct their password. Moreover, this method aims at the generation effect.

Evaluation. Yan has conducted a field trial on the memorability of passphrases [62] in two sessions at the start and the beginning over a time period of three months. The participants were told to generate a password for their students university account from a pass phrase. Before password creation, participants were given advice on how to construct a memorable and secure password from a pass sentence.
The participants were all students.

Analysis of the evaluation. As password memorability was observed over three months during the whole academic term in a field trial, Yans user study takes place in a more realistic environment than the laboratory and involves everyday memory. Password authentication is consequently not evaluated as a primary, but rather as a secondary task. We also assume that students have logged in several times during the term, so that the study also exploits the spacing effect. Moreover, since participants had to create their passwords at the beginning of the study, the generation effect is addressed as well.

3.2.3 PassShapes

Design. The PassShape system, shown in Figure 1d, is based on the observation that we often use geometric shapes as an aid when we enter PIN numbers. A PassShape therefore consists of stroke-based geometric forms drawn by hand, whereby each stroke is tantamount to a certain number. The stroke order is fixed [59].

Analysis of the design. Of the graphical password systems we consider in this paper, the drawmetric PassShapes system is the only system that addresses not only the pictorial superiority effect, but also our procedural memory: Through the use of drawings, the PassShape is processed and stored as a motor scheme, that is, as a particular movement sequence. Moreover, research has demonstrated that a fixed stroke order improves the memorability of characters [25], a fact that also helps with the memorability of geometric shapes [59].

Evaluation. Memorability evaluation of the PassShape system has been done by Weiss and De Luca. The user study took place over a period of ten days with a session on the first day and another two sessions on the fifth and tenth day respectively. The memorability of the system was studied in three groups, of which two groups were using PassShapes for authentication. During the first session, participants of these two groups were each given a system-generated PassShape first. Participants in one group were instructed to memorize the PassShape in correct stroke order without further advice, participants in the other group were told to correctly draw their PassShape 24 times. After this learning phase, all participants had to answer a questionnaire, which also served to distract short-time memory. During the following retrieval phase, the participants were instructed to repeat their PassShape once more. Retrieval was then again tested in the second and third session after five and ten days respectively. Participants of the first PassShape group who failed to successfully repeat their PassShape were presented with the correct PassShape once more. Participants of the second PassShape group had to repeat their PassShape for another 24 times if they could not remember it correctly.
The participants of the user study were all part of the university environment. The proportion of male and female participants was balanced.

Analysis of the evaluation. Memorization in the first PassShape group refers to the pictorial superiority effect alone, while learning the PassShape in the second group also trains procedural memory through repetitive rehearsal. Conducting several sessions over ten days, the

	Proactive password checking	Passphrases	PassShapes	PassPoints	PassFaces	Cognitive passwords
Feedback effect					X	
Spacing effect						
Pictorial superiority effect			X	X	X	
Testing effect						
Von Restorff effect					X	
Generation effect	X	X			X	
Self-reference effect						X
Memory interference						
Mnemonics		X		X		
Everyday memory						
Procedural memory			X			
Cued recall or recognition				X	X	X

Table 1: Aspects of memory in the *design* of current knowledge-based authentication systems.

	Proactive password checking	Passphrases	PassShapes	PassPoints	PassFaces	Cognitive passwords
Feedback effect				X		
Spacing effect	X	X	X	X	X	X
Pictorial superiority effect			X			
Testing effect	X		X	X		
Von Restorff effect						
Generation effect	X	X		X		
Self-reference effect						
Memory interference	X					
Mnemonics	X		X			
Everyday memory		X			X	
Procedural memory			X			
Cued recall or recognition						

Table 2: Aspects of memory in the *evaluation* of current knowledge-based authentication systems.

study aims at involving the spacing effect and distributed practice, so that information is more likely to be stored in long-term memory. As PassShapes are geometric forms and therefore more abstract than pictures, results of the user study have also shown that participants tried to process their PassShape more deeply by using mnemonic techniques, such as connecting the appearance of a shape to objects that looked similar. The user study also exploits the testing effect as users had to reenter their PassShape after a short break.

3.2.4 PassPoints

Design. Passwords of the PassPoints system are given by a sequence of points chosen by the user in an image (see Fig. 1e). During authentication, users have to click on these points in the right order in a certain range of tolerance [60].

Analysis of the design. As PassPoints is a locimetric graphical password system, the password design exploits the pictorial superiority effect. It also refers to a mnemonic technique called the "method of loci" where information is linked to certain landmarks [51]. Moreover, password input is based on cued recall [60].

Evaluation. Wiedenbeck et al. tested memorability of PassPoints in a user study. Evaluation of PassPoints has been done over multiple sessions over a period of six weeks, consisting in the three phases of 1) password creation, 2) password learning and 3) password retention. The sessions occurred in week 1, week 2, and week 6 respectively. In week 1, participants were first told to create a password of five points and then to learn it by entering it repeatedly. After password creation, the entire password was displayed along with graphical highlighting of the range of tolerance around each point and numbers showing the input order. During the learning phase, the participants had to enter their passwords repeatedly until they achieved ten successful inputs, and were given binary feedback on the input correctness as well as on the total count of correct and incorrect entries. After learning the password, the participants had to fill out a questionnaire which also served as a distraction task between the learning and first retention phase. During the second and third session in week 2 and week 6, password retention was studied again. During each retention phase, the participants had only to enter their password correctly one time. Testing was done individually in a laboratory. The participants were all part of the university environment and the proportion of male and female participants was balanced.

Analysis of the evaluation. The evaluation of PassPoints over six weeks during several sessions clearly aims at the spacing effect and distributed practice, which helps to transfer information from short-term to long-term memory through rehearsal. Feedback is given in order to improve learning. Moreover, the password creation phase refers to the generation effect, according to which the users will create retrieval cues on his own and thus process the information more deeply. This user study exploits the testing effect as well since participants had to retrieve their passwords from memory after a short break.

3.2.5 PassFaces

Design. PassFaces passwords consist of a sequence of pictures of human faces, called passfaces. During authentication, users have to pick their passfaces from a set of human faces, containing both the passfaces and randomly chosen distractor faces. An example passfaces grid is shown in Figure 1f. Users are familiarized with their passfaces in an enrollment procedure implemented in the system that consists of an introduction and a learning phase. During the introduction phase, the users choose their passfaces and each of the passfaces are shown individually to the users. During the learning phase, the users have to practice their passfaces until they have successfully identified them at least four times. When practicing the passfaces during the first rounds, the passfaces are highlighted from the distraction faces and users are also provided with immediate feedback whenever they choose incorrect faces [4].

Analysis of the design. PassFaces is a cognometric graphical system based on the recognition of faces: It thus takes advantage of the pictorial superiority effect and of the fact that we remember information more easily in recognition than in recall tasks. Passfaces are memorized by exploiting the generation effect when the users select their passfaces, the Von Restorff effect during the first input round, by rehearsing them repeatedly and by giving feedback on incorrect input.

Evaluation. Evaluation of the PassFaces system has been done by Valentine [55] over the time period of one month. After all participants had gone through the PassFaces enrollment procedure, they were assigned to different groups with different login rates respectively. The first group had to log in on every working day for two weeks, the second group was told to log in one week after enrolement and the third group was instructed to do so after one month of password creation.
All participants were part of the university environment. Moreover, Brostoff and Sasse have conducted a field trial [13] during an academic one-term course, where the participants (all students) had to login in order to be able to do the coursework. During the course, participants had to use their passfaces repeatedly.

Analysis of the evaluation. The user studies for the PassFaces system all take advantage of the spacing effect. While Valentine's study took place in the laboratory, Brostoff and Sasse have tested memorability for PassFaces under more realistic conditions, involving everyday memory and testing password authentication as a secondary and not as a primary task.

3.2.6 Cognitive passwords

Design. Cognitive passwords are set up as a quiz in which the correct answers to several questions grant access to the system (see Fig. 1c. The users' answers thus represent several passwords they have to enter (instead of only one) that are based on the users' personal history, experiences and interests. Examples for possible questions are "What is your mother's maiden name?", "What was the name of your first pet?" etc. [63].

Analysis of the design. Cognitive password systems exploit the self-reference effect which implies that we remember information better once we refer it to ourselves as deep processing and the creation of personal retrieval cues are encouraged. Moreover, cognitive passwords use cued recall to retrieve information from memory.

Evaluation. The cognitive password system and its memorability were evaluated in a user study conducted by Zviran and Haga over a time period of three months during two sessions, one at the beginning of the user study and one at the end of the third month. In the first session, participants were presented with a set of questions on paper and were told to provide the answers, that is, their passwords. After three months, participants had to follow the same procedure again. The answers of the first session where then compared with the answers of the second session.
All participants were students, the main part (85 percent) was male [63].

Analysis of the evaluation. The study does not consider any of the discussed memory effects or memory characteristics except for the spacing effect.

4 Conclusion and discussion

In this paper, we have addressed the topics of human memory and memorability in the context of the design and the evaluation of current knowledge-based authentication systems. We have given a detailed overview of the findings of cognitive psychology and have explained how information processing in human memory works. We then have presented several interesting aspects of memory, such as the pictorial superiority effect, the spacing effect or the interference of multiple items in memory, which have been demonstrated to influence information processing. Based on these insights, we have first suggested how these psychological findings could be applied to usable security research. We then have analyzed whether current knowledge-based authentication systems adequately take into account the different aspects of memory in their design and evaluation. We find that all considered systems include memorability as a usability criterion and take into consideration psychological research. In the systems' design, the memory effects that are used the most are the pictorial superiority effect and the generation effect. Moreover, cued recall and recognition rather than free recall are addressed. In the evaluation of memorability, the spacing effect is the most prevailing memory effect, followed by the testing and the generation effect.

Current usable security research has thus realized the need for psychological insights in order to build and evaluate knowledge-based authentication systems that are more memorable than traditional alphanumeric passwords. However, our analysis suggests that especially the design of user studies in order to evaluate memorability is still a vague topic. Although several aspects of memory are considered during user studies, we could not find best practices: For instance, for the systems we presented, the period of time for the spacing effect ranged from one week to a whole academic term and the sessions were arbitrarily spread over this timespan. Also, user studies are always conducted without the help of psychologists.

Moreover, field trials in order to evaluate everyday memory are still sparse, and a laboratory-based study has several shortcomings. Password authentication is often tested as a primary task and participants thus may pay more attention than they would in real life, a factor that could influence password retention. The laboratory situation could affect context-dependent memory, too, so that users have less retrieval difficulties than they would have in everyday life where the context changes often. Also, multiple passwords and possible memory interference are seldomly evaluated.

Another issue is the test group: In all of the user studies we examined and even in the field trials, the participants form a rather homogeneous group. In most cases, they are students or faculty members with a high level of education. It thus has to be considered whether the homogeneity of the test group affects the evaluation results. Also, even though in most of the presented studies, the proportion of male and female participants was almost balanced, it should be kept in mind that cognitive research has shown differences between male and female information processing and memory [34].

Finally, there are other interesting aspects of memory and memorability we did not include in our explanations, for instance people's specific learning styles when processing new information. Issues like this could be considered concerning the design and evaluation of memorable passwords and need further study.

References

[1] https://discussions.apple.com/thread/5109842 [online; accessed 30-june-2014].

[2] https://www.microsoft.com/security/pc-security/password-checker.aspx [online; accessed 30-june-2014].

[3] http://www.realuser.com/enterprise/news/logo_and_graphics.htm [online; accessed 30-june-2014].

[4] The science behind passfaces. Real User Corporation, 2002.

[5] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):41–46, 1999.

[6] R. C. Atkinson and R. M. Shiffrin. Human memory: A proposed system and its control processes. *The psychology of learning and motivation*, 2:89–195, 1968.

[7] A. D. Baddeley. The psychology of memory. *The essential handbook of memory disorders for clinicians*, pages 1–13, 2004.

[8] A. D. Baddeley and G. J. Hitch. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.

[9] H. P. Bahrick and E. Phelphs. Retention of spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(2):344, 1987.

[10] M. Bishop and D. V Klein. Improving system security via proactive password checking. *Computers & Security*, 14(3):233–249, 1995.

[11] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

[12] T. Brashers-Krug, R. Shadmehr, and E. Bizzi. Consolidation in human motor memory. *Nature*, 382(6588):252–255, 1996.

[13] S. Brostoff and M. A. Sasse. Are passfaces more usable than passwords? a field trial investigation. In *People and Computers XIVUsability or Else!*, pages 405–424. Springer, 2000.

[14] A. S. Brown, E. Bracken, S. Zoccoli, and K. Douglas. Generating and remembering passwords. *Applied Cognitive Psychology*, 18(6):641–651, 2004.

[15] S. K. Carpenter, H. Pashler, J. T. Wixted, and E. Vul. The effects of tests on learning and forgetting. *Memory & Cognition*, 36(2):438–448, 2008.

[16] N. J. Cepeda, N. Coburn, D. Rohrer, J. T. Wixted, M. C. Mozer, and H. Pashler. Optimizing distributed practice. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, 56(4):236–246, 2009.

[17] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3):354, 2006.

[18] F. I. Craik and E. Tulving. Depth of processing and the retention of words in episodic memory. *Journal of experimental Psychology: general*, 104(3):268, 1975.

[19] F. I. M. Craik and R. S. Lockhart. Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behaviour*, 11:671–684, 1972.

[20] A. De Angeli, L. Coventry, G. Johnson, and K. Renaud. Is a picture really worth a thousand words? exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies*, 63(1):128–152, 2005.

[21] R. Dhamija and A. Perrig. Deja vu: A user study using images for authentication. In *Proceedings of the 9th USENIX Security Symposium*, pages 45–58. Usenix Denver, CO, 2000.

[22] V. S. R. Douglas L. Nelson and J. R. Walling. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory*, 2(5):523–528, 1975.

[23] M. Fabiani and E. Donchin. Encoding processes and memory organization: a model of the von restorff effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1):224, 1995.

[24] S. J. Gillihan and M. J. Farah. Is self special? a critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological bulletin*, 131(1):76, 2005.

[25] F. B. d. Giovanni. Order of strokes writing as a cue for retrieval in reading chinese characters. *European Journal of Cognitive Psychology*, 6(4):337–355, 1994.

[26] A. M. Glenberg. Influences of retrieval processes on the spacing effect in free recall. *Journal of Experimental Psychology: Human Learning and Memory*, 3(3):282, 1977.

[27] D. Godden and A. Baddeley. When does context influence recognition memory? *British Journal of Psychology*, 71(1):99–104, 1980.

[28] L. Grossman and M. Eagle. Synonymity, antonymity, and association in false recognition responses. *Journal of Experimental Psychology*, 83(2p1):244, 1970.

[29] F. Haist, A. P. Shimamura, and L. R. Squire. On the relationship between recall and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(4):691, 1992.

[30] H. L. Hollingworth. Characteristic differences between recall and recognition. *The American Journal of Psychology*, pages 532–544, 1913.

[31] B. Ives, K. R. Walsh, and H. Schneider. The domino effect of password reuse. *Communications of the ACM*, 47(4):75–78, 2004.

[32] H. Kay. Learning and retaining verbal material. *British Journal of Psychology*, 46(2):81–100, 1955.

[33] D. V. Klein. Foiling the cracker: A survey of, and improvements to, password security. In *Proceedings of the 2nd USENIX Security Workshop*, pages 5–14, 1990.

[34] M. W. Matlin. *Cognitive Psychology*. John Wiley & Sons, 8th edition international student version edition edition, 2014.

[35] M. A. McDaniel, H. L. Roediger, and K. B. McDermott. Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, 14(2):200–206, 2007.

[36] G. A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[37] A. Miyake and P. Shah. *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press, 1999.

[38] I. Neath. *Human memory: An introduction to research, data, and theory.* Thomson Brooks/Cole Publishing Co, 1998.

[39] T. O. Nelson and R. Rothbart. Acoustic savings for items forgotten from long-term memory. *Journal of Experimental Psychology*, 93(2):357, 1972.

[40] J. Nielsen. *Usability engineering*. Morgan Kaufmann, 1st edition, 1993.

[41] A. Paivio and K. Csapo. Picture superiority in free recall: Imagery or dual coding? *Cognitive psychology*, 5(2):176–206, 1973.

[42] H. Pashler, D. Rohrer, N. J. Cepeda, and S. K. Carpenter. Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, 14(2):187–193, 2007.

[43] H. L. Roediger. Memory metaphors in cognitive psychology. *Memory & Cognition*, 8(3):231–246, 1980.

[44] H. L. Roediger III, A. L. Putnam, and M. A. Smith. 1 ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation-Advances in Research and Theory*, 55:1, 2011.

[45] D. Rohrer, K. Taylor, H. Pashler, J. T. Wixted, and N. J. Cepeda. The effect of overlearning on long-term retention. *Applied Cognitive Psychology*, 19(3):361–374, 2005.

[46] D. L. Schacter. The seven sins of memory: Insights from psychology and cognitive neuroscience. *American psychologist*, 54(3):182, 1999.

[47] N. J. Slamecka and P. Graf. The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, 4(6):592, 1978.

[48] L. R. Squire. Mechanisms of memory. *Science*, 232(4758):1612–1619, 1986.

[49] L. Standing. Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, 25(2):207–222, 1973.

[50] L. Standing, J. Conezio, and R. N. Haber. Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, 19(2):73–74, 1970.

[51] R. J. Sternberg, K. Sternberg, and J. Mio. *Cognitive Psychology*. Thomson/Wadsworth, 6th edition, 2011.

[52] E. Tulving. The effect of alphabetical subjective organization on memorizing unrelated words. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 16(3):185, 1962.

[53] E. Tulving and Z. Pearlstone. Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, 5(4):381–391, 1966.

[54] E. Tulving and D. M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352, 1973.

[55] T. Valentine. An evaluation of the passface personal authentication system. Technical report, Goldmsiths College University of London, 1998.

[56] H. Von Restorff. Über die wirkung von bereichsbildungen im spurenfeld. *Psychologische Forschung*, 18(1):299–342, 1933.

[57] K.-P. L. Vu, R. W. Proctor, A. Bhargav-Spantzel, B.-L. Tai, J. Cook, and E. E. Schultz. Improving password security and memorability to protect personal and organizational information. *Int. J. Human-Computer Studies*, 65:744–757, 2007.

[58] H. S. E. Waters and W. E. Schneider. *Metacognition, strategy use, and instruction.* Guilford Press, 2010.

[59] R. Weiss and A. De Luca. Passshapes: utilizing stroke based authentication to increase password memorability. In *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, pages 383–392. ACM, 2008.

[60] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon. Passpoints: Design and longitudinal evaluation of a graphical password system. *Int. J. Human-Computer Studies*, 63:102–127, 2005.

[61] J. B. Worthen and R. R. Hunt. *Mnemonology: Mnemonics for the 21st century*. Psychology Press, 2011.

[62] J. J. Yan, A. F. Blackwell, R. J. Anderson, and A. Grant. Password memorability and security: Empirical results. *IEEE Security & privacy*, 2(5):25–31, 2004.

[63] M. Zviran and W. J. Haga. A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal*, 36(3):227–237, 1993.

How to Design Security Questions

Christian Becker

Abstract— Passwords nowadays are used to authenticate users on websites, devices and services in general. To ensure a minimum level of security it is recommended to use complex and unique passwords. This can lead users to forget them and if so, password recovery mechanisms are required. Security or challenge-response questions represent a common recovery mechanism. At registration time a user answers a set of self-defined or pre-defined questions. The answers will be stored by the corresponding service or device. To recover a password the user is prompted to answer a subset of the specified questions. However, if not designed properly questions can be vulnerable to attacks like guessing or observing. Furthermore usability issues related to memorability or applicability can arise depending on the type of question used. Hence, in this paper common problem areas of security questions will be identified and criteria introduced to design both secure and usable questions. The paper will conclude with alternative question-based approaches and a brief outlook.

Index Terms—Security Questions, Challenge-Response Questions, Fallback-Authentication, Usability

1 Introduction

In the last decades the use of passwords changed dramatically. Originally they were used to authenticate users operating locally on a terminal. As network technology began to evolve and computers became more ubiquitous, passwords became also essential for connecting to remote accounts [27]. Nowadays continuously new services and websites arise demanding for new and unique passwords. To avoid compromise, passwords should have a certain length and degree of complexity. However, this can lead users to forget their passwords and loose access to their accounts [7, 18]. In this case a second authentication mechanism, called fallback authentication, is required to re-authenticate the user.

Security or challenge-response questions represent one kind of fallback authentication. In theory they can also be used as an additional security level to complement existing mechanisms like passwords [17] or as a primary authentication mechanism [15]. However as described by Just [15] this implicates a variety of disadvantages. For example, to improve security many question-based authentication systems require more than one question to be answered before access is granted. This can lead to a longer-lasting authentication procedure which might not be tolerable for every user. The authentication of the user might further be delayed by an additional "out-of-bound" authentication step, for example if an authentication code is sent via email. Therefore this paper focuses on the use as fallback authentication.

Security questions were originally developed as an alternative authentication mechanism to conventional passwords. According to Zviran and Haga [29] this kind of questions has a better recall rate than passwords, but they may be easily answered by guessing or observing. In the past this has been exploited in practice. For instance, in 2008 an attacker could get access to Sarah Palin's (former Republican vice presidential candidate) email account by answering security questions referring to her date of birth, zip code and place/date when she first met her husband [4]. This is a problem especially nowadays due to data being publicly available, e.g. in social networks [23]. Gupta et al. [11] showed that data available through public Facebook pages can be used to derive a users' interests. If these interests are known authentication systems referring to users' preferences can be broken more easily.

- *Christian Becker is studying Media Informatics at the University of Munich, Germany, E-mail: christian.becker@campus.lmu.de*
- *This research paper was written for the Media Informatics Advanced Seminar 'Special Aspects of Usability', 2014*

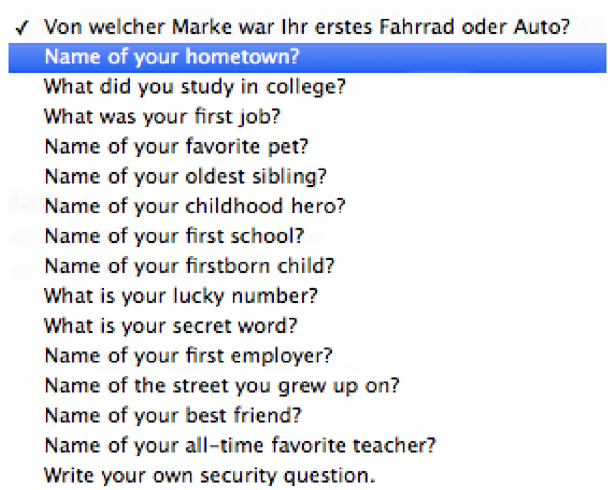


Fig. 1. Screenshot: The range of questions offered by Apple. In this case a question can be chosen at registration or afterwards [2].

In the next chapter the process of question-based fallback-authentication will be explained and examples for questions will be given. The subsequent chapters will deal with the existing types of questions, their general issues and design criteria to enhance security and usability of corresponding question systems. Afterwards the security and usability of the question types will be analysed.

2 What are Security Questions?

Security questions are commonly used by companies like Apple, Facebook or Yahoo [1, 6, 28] as fallback authentication to help users recover their passwords after those have been forgotten or compromised.

Question-based authentication systems consist of question-answer pairs which are defined by the service, user or both. For example, Figure 1 shows the range of questions offered by Apple. The first one at the top was defined and selected by the user.
At account registration or afterwards a user is prompted to answer a certain amount of questions. These questions can be defined by the corresponding service (pre-defined) or by the user (user-defined). If questions are pre-defined, in most cases the user can choose which questions to answer from a catalogue of questions [23]. Once answered, questions and answers are stored by the corresponding service. In case fallback authentication is needed all or a subset of the questions answered at registration are presented to the user. If all or

some of answers are correct access is granted.

Security questions can differ in a variety of aspects, for example in the amount of questions that need to be answered correctly before access is granted, the number of chances a user is given or the types of questions used. These aspects will be evaluated in chapters 3 and 4.

Based on the work of Wood [27] and Just [15], authentication systems can be classified into knowledge-based (e.g. passwords or personal identification numbers), object-based (e.g. tokens or smartcards) or id-based (e.g. fingerprints or iris-scans). Challenge-response systems in general also are counted among knowledge-based systems. This implies several advantages. For example, information cannot get physically lost or stolen as it is possible when using object-based systems. It is easier to replace information (security questions can be changed) as in case of using id-based systems [21]. Knowledge-based authentication systems result in lower error rates than id-based systems due to the fact that there is no error rate in measurement and measurement does not depend on the environment [21]. Furthermore it is impossible to create a physical spoof of a template [5]. However information to answer security questions or the answers themselves can be compromised by attackers, for example by using information that is publicly available. Corresponding information can also be forgotten by users. Therefore usability, in particular memorability and repeatability are also critical factors. This will be evaluated in the following chapters.

3 Types of Security Questions

As mentioned in the previous chapter security questions amongst others can differ in their question type. For instance, as well as personal or cognitive questions, sensitive questions determine the kind of information they refer to. They ask about sensitive personal information like ATM pin codes or Social Security Numbers. As mentioned by Rabkin [23] and Garfinkel [8] this information can be compromised. Therefore in this paper the focus will be on personal respectively cognitive questions. Their advantages and disadvantages will be overviewed.

3.1 Personal or Cognitive

In contrast to sensitive questions, personal or cognitive questions deal with a users' personal environment, history or background. They can be defined by the user or institution/service provider. In general answers can be learned or guessed by attackers [23].

3.2 Fact-based and Opinion-based

Fact-based and opinion-based questions also refer to the question's content rather than to its formal aspects. Besides personal or cognitive questions they represent another approach to classify the kind of information being asked for. That is why in practice fact-based and opinion-based respectively personal/cognitive questions can be used in combination.
The first ones, fact-based questions, are related to statements by an individual containing facts [15]. It is assumed that the corresponding answer varies only little over time. Therefore, as stated by Just [15] it may be easier for an attacker to learn the answer. The basis of the latter question type, opinion-based questions, is said to be the beliefs of an individual. As beliefs change over time the answer might change more over time than in case of fact-based questions. Yet these should be less pervasive since individuals need to deal with their own opinions less frequently in their daily lifes [15].

3.3 Open

In contrast to the first three question types the following one determines the formal aspects of the question.
Open questions and their corresponding answers are defined by users and hence are not limited to a special kind of information in their reference. These questions and answers are assumed to result in improved memorability and applicability. Furthermore users can choose questions with corresponding answers more difficult to find in public sources. However user-defined questions may result in insecure questions [16].

3.4 Fixed

When this type of question is used, questions and/or possible answers are defined by the service provider or institution the authentication mechanism is offered by. To increase applicability users often can choose from a list of predefined questions. Thereby they are prevented from specifying possible insecure questions. However, memorability and repeatability may suffer [14][15].

3.5 Controlled

Controlled questions are fixed except for one part, e.g. a word. A question could look like: ”What is Name's middle name?”. To complete the question a user not only needs to provide the corresponding answer but also a value for the missing part. In the example a value for ”Name” needs to be provided. Additionaly to increase memorability and/or repeatability a hint could be stored at registration. Though it must be chosen carefully to avoid providing too much information to an attacker [15].

4 Designing Secure and Usable Security Questions

As mentioned above question types will be evaluated with regard to security and usability in chapter 7. However to support a design of secure and usable questions a variety of design criteria have been described in related work. These criteria will be addressed in the following section. Furthermore common issues of security questions and approaches to solve them will be evaluated in this chapter.

4.1 Privacy

First of all one should consider that security questions can lead users to provide sensitive information. To protect the users' privacy this information needs to be handled carefully by the corresponding service. For example, it should collect as less information as possible but as many as necessary to provide a working recovery mechanism. The information collected should not be used for anything else than for recovery and, if possible, users should be able to decide which information they want to provide. Furthermore answers should be hashed before being stored. If a question system uses multiple questions a single hash should be computed for all answers. This ensures that an attacker needs to guess all answers before he is able to determine if he was successful [15].

Finally one should mention that security questions are vulnerable to shoulder surfing. That is why, as in case of passwords, the display of the answers should be obscured. Although this is possible it is often avoided since answers can contain varying capitalization and punctuation [14].

4.2 Security

Security is another crucial factor when designing security questions. Therefore this section will address corresponding criteria and aspects to support a secure design.

4.2.1 Guessing and observing

In general questions should resist attacks based on guessing and observing. The more difficult it is to use these kinds of attacks successfully the more secure a question is. For instance, questions asking for mothers' maiden names are not secure since information referring to birth and marriage are publicly available in parts of the US and can be used in observing attacks to determine the maiden name [9]. As mentioned by Just [17] and Rabkin [23] especially information to answer security questions can be publicly available since people are used to share information on social networks. To avoid compromise, questions should not ask for that kind of information. Furthermore, all answers should be validated before an indication of success or failure is given. Otherwhise, attackers can guess the answer to one question at a time [15].

4.2.2 Amount of questions

Moreover the relation between question and answer as well as the amount of questions are important factors. To improve security questions and answers should be independent from each other [16]. Thereby deriving the answer from the question can be avoided. According to Just [15] the amount of questions can cause both positive and negative effects on security and usability. If fewer questions are asked, it can lessen the recall requirements and result in fewer repeatability mistakes. Hence usability increases. However if too few questions are asked security can suffer. Usually the amount of questions depends on the attack that should be prevented [16].

4.3 Usability

Besides security usability is another important aspect in designing securiy questions. In the following subsections criteria affecting usability will be introduced.

4.3.1 Applicability, Recall and Formatting

According to Just [15] the usability of question-based authentication mechanisms can be evaluated using three critical factors. First of all questions should be applicable to as many users as possible. Questions like 'What was the name of your first pet?' are not applicable to everyone since not everyone has been owning a pet. Secondly security questions should be designed to such an extent that users are able to recall the corresponding answer at least in some way. Finally in the best case users should be able to recall the exact answer. This however, cannot be gueranteed since certain questions like 'What is the name of the street you grew up in?' require a certain formatting or notation. In the example 'street' can be abbreviated. If a user assumes a different notation it can result in a wrong answer. Therefore in practice only questions with fixed formatting should be used.

4.3.2 First-Time Events and Favorites

Furthermore questions referring to "first-time events" should be avoided since these are more difficult to recall for older people [14]. Rabkin [23] as well as Bonneau et al. [3] also found that questions asking for favorites are inapplicable since preferences are easily guessable by using public sources, e.g. social networks. Often they offer the chance to track favorite books, music or movies. Also public Facebook pages can be used to determine a users' interests and thereby preferences [14].

5 How to attack Security Questions

To achieve the desired level of security when designing security questions it needs to be known how these questions can be attacked. Thus this chapter will give an overview about common techniques.

5.1 Guessing and automated Guessing

Guessing answers can be based on answer entropy or an statistical analysis of the question and its most common answers. It can be broken down to: Blind guess and focused guess. Blind guess considers not only all possible answers based on the answer alphabet, but also letter distribution and common answer sets [25]. In contrast to blind guess, focused guess considers the question itself to such a degree as it is used to determine the distribution of likely answers [16].

5.2 Observing

Observing is a more sophisticated attack. It relies on gathering publicly available information about the user, his environment or preferences. For instance, this information can be obtained by using social networks. Possible answers are then derived based on this information [23].

6 Measuring the Level of Security

To determine if a question is secure its level of security needs to be measured. This can be achieved by measuring the entropy and false positive/false negative rate. This will be outlined in the following sections. However before, ethical concerns that may arise in studies when users are prompted to not only provide their questions but also the corresponding answers need to be addressed.

6.1 General Issues

As mentioned before when users are not only asked to provide their questions but also answers to their questions ethical concerns may arise because they may fear that this information may be used to compromise their accounts. Hence fictious answers may be provided and therefore reliable results cannot be guaranteed. Though according to Just and Aspinall[16] reliability can be increased in studies by not asking users for their answers but only for their questions.

6.2 Entropy

In addition to the difficulty of guessing and observing, the security of questions also can be determined by measuring the entropy of answers and the false positive/false negative rate.
The amount of entropy defines the amount of information an information source contains. The more information it contains, the more difficult it is to reproduce the same amount of information within a new information source [25], e.g. an answer to a security question.
In general the entropy of answers should be similar to that for routine authentication [15]. However this can be difficult since answers to security questions are often based on words included in a dictionary. As mentioned by Just [14] most dictionaries have between 2^{16} and 2^{20} words. Compared to that an eight-character password only made up of lowercase characters, ten numbers and 32 punctuation characters has 2^{38} possibilities. Therefore at least two questions would be needed to reach the same security level of an eight-character password.

6.3 False Positives/false Negatives

By using preference-based question systems people can make mistakes since preferences can change over time. Hence a threshold must be introduced to distinguish between legitimate users and attackers. This threshold must be chosen to such an extent that the false negative rate is minimized (likelihood to reject legitimate users decreases) and the false positive rate is maximized (access to illegimate users should not be granted) [13].

7 Analysis

In the following sections different types of security questions will be analyzed in terms of security and usability. Due to the fact that open and personal/cognitive questions have been evaluated more closely in studies, focus will be on these question types.

7.1 Open Questions

Just [16] conducted a user study to evaluate usability as well as security in terms of open questions. This study consisted of two stages. In the first stage each participant was asked to come up with three questions and corresponding answers he or she would use for password recovery. The questions were submitted to an online questionnaire. The answers were kept private by the participants to avoid a fear of being misused/ethical concerns and hence to improve reliability. Overall 94 students submitted their questions and answers. A few weeks later all participants were asked to recall their answers in the second stage.

The submissions have been evaluated with regard to three common attacks: blind guess, focused guess and observation (see chapter 5 for more information about the different kinds of attacks). For each of the attacks three security levels (low, medium and high) were specified. However the meaning of the levels differed. In case of blind guess and focused guess the levels were derived from comparable levels

Table 1. Rating by security level for different attack types and questions [16]

Attack type	Low	Medium	High	Total
Blind Guess	174	4	2	180
Focused Guess	167	0	13	180
Observation	124	54	2	180

Table 2. Rating by security level for different attack types and users [16]

Attack type	Low	Medium	High	Total
Blind Guess	5	19	36	60
Focused Guess	5	28	27	60
Observation	24	34	2	60

for passwords and were related to the amount of possible answers or entropy. The security level of a question was considered to be low if the amount of possible answers was not exceeding 2^{34} answers. To reach a medium security level between 2^{34} and 2^{48} answers were necessary. High was assigned if there were more than 2^{48} possible answers. In case of observation the security level referred to the amount of people the answer was known to. For example, low was given if the answer was publicly available and high if the answer is neither known to family and/or friends nor believed to be publicly available.

Table 1 indicates that answers to generated questions by participants of the study are not sufficiently secure and hence are vulnerable to attacks. For blind guess about 96%, for focused guess about 92% and for observation about 68% of the answers were considered to have a low level of security or entropy. If multiple questions were combined the overall security level per user could be improved (see Table 2). However the amount of questions needed to gain a certain level of security depended on the kind of attack. For example, in case of blind guess 36 users reached security level "high" whereas for focused guess only 27 reached "high" for the same amount of three questions. Hence for focused guess more questions are needed to reach the same level of security. When all three attack types were applied simultaneously no user achieved "high" for every attack and 31 users did not achieve any "low" rating. Of the remaining users 15 achieved a "high/medium/medium" rating and twelve a "high/high/medium" rating. According to Just [16] there was no perfect set of questions but by over half of the participants a moderate level of security could be achieved. In practice it is recommended to filter or reject questions that are not sufficient secure.

In the study participants were also asked for their security perception. 163 questions were considered by participants as to be "very difficult" or "somewhat difficult" to be answered by strangers. When the security rating is kept in mind this shows that users tend to overestimate the security of their questions. Besides sharing information with family or friends is common since 98 questions were considered as to be "not difficult at all" to be answered by friends or family.

Memorability and repeatability were also important aspects in the context of the study. Although the participants were given the chance to define their own (open) questions, difficulties in recalling persisted. After error compensation 7 of 60 participants could not recall at least one answer exactly after approximately 23 days. Responses and comments by four of them indicated that it was an issue of repeatability. One possible reason could be that 40% of the questions referred to past or first-time events. As mentioned before answers to these kind of questions can be difficult to remember for older people or can change over time. Furthermore some participants may have given fictional answers they could not remember [16].

7.2 Personal or Cognitive Questions

Zviran and Haga [12, 30] evaluated personal or cognitive questions in terms of memorability and compared them to passphrases as well as user-generated, system-generated and associated passwords. Overall 103 graduate students took part in a study. In a questionnaire each participant was asked to create a password of up to eight alphanumeric characters as well as a passphrase of up to 80 alphanumeric characters. Furthermore each of the 55 participants was assigned a system-generated random alphanumeric password. The remaining 48 participants were assigned a pronouncable password. Besides the passwords and the passphrase each participant was prompted to come up with, a list of 20 word associations and a set of 20 predefined questions, whereas six of them were fact-based and 14 opinion-based questions, should be answered. Three months later the participants were asked to recall the corresponding answers, passwords and the passphrase.

Personal or cognitive questions could achieve the highest recall rate in the study. On average 74% of all 20 questions could be recalled. However only 2 of 103 participants could recall all answers. When fact-based and opinion-based questions were evaluated separately for fact-based questions the average recall rate was 83.7%. In case of opinion-based questions on average 74.8% of the questions could be recalled. The average recall rates for the remaining password mechanisms were as follows: associative passwords: 69%, self-generated passwords: 27.2%, system-generated passwords: 24% (pronouncable: 38%, random alphanumeric: 24%) and passphrases: 21.4%. As mentioned by Zviran and Haga [12, 30] this indicates that, at least in terms of memorability, personal or cognitive questions are superior to common passwords.

To determine in what way the mechanisms are secure, Zviran and Haga [12] also evaluated the guessability by significant others (family members, spouses and friends) for personal or cognitive questions and associative passwords. On average 38% of all cognitive items could be guessed correctly. For fact-based questions the average guessing rate on the average was 44.6%, for opinion-based questions 32.5%. Family members guessed the most items correctly (60%). The guessing rate for spouses was 41% and for friends 23.5%. Hence by increasing the social distance the guessing rate drops. Compared to associative passwords the guessing rate for personal or cognitive questions was higher. Before having been given associative cues the significant-others on the average could answer 25.5% of the questions. When the theme was given the answer rate increased to 45%. This supports the statement that the theme needs to be kept private.

Podd et al. [22] compared conventional passwords to word association and cognitive security questions. 86 students were asked to answer a questionnaire containing 40 "cognitive questions" whereas 20 of them were opinion-based and 20 fact-based. Additionaly they had to respond to 20 cue words within a word association test. Finally every participant was assigned an eight-character, alphanumeric password of the form three-letter word, one digit, four-letter word and prompted to come up with an eight-character, alphanumeric password by herself or himself. Two weeks later the participants took part in a recall test. In the course of this test each participant was also asked to specifiy a significant-other for guessing purposes. On average 80% of the cognitive questions could be recalled correctly. Thereby compared to the other mechanisms cognitive questions could achieve the highest recall rate. However at 39.5% guessing rate cognitive questions could be guessed more easily than associated or conventional passwords. When it was distinguished between opinion-based and fact-based questions the latter produced a very high correct guessing rate at 56%. Compared to conventional passwords and word association it was the highest guessing rate. However on average 88% of the answers could be recalled. For opinion-based questions recall rate was at 72%, guessing rate at 23%. Podd et al. [22] concluded that the best of the opinion-based items could achieve acceptable rates both of guessing and recall. For example, the following questions were among the best five opinion-based items: "Who is your favorite school teacher?" (Recall rate: 77%, guessing rate: 15%), "What is your favorite sandwich filling?" (Recall rate: 70%, guessing rate: 20%) or "Who is your favorite actor?" (Recall rate: 70%, guessing rate: 13%). According to Podd et al. [22] in practice a set of five cognitive items in combination with a conventional password could improve security.

Further analysis of security questions used by online services was conducted by Rabkin [23] as well as Schechter et al. [24].
In his work Rabkin [23] evaluated 20 banking websites in terms of fallback authentication. 15 of these websites used security questions, either sensitive or personal/cognitive, to recover a users' passwords. At registration the user was shown at least one subset of questions derived from a overall pool of questions. He or she was asked to choose minimum one question from each subset and answer it. Although this procedure was the same across all websites the circumstances were not. For example, the size of the overall question pool as well as the questions needed to be answered varied. When the security of the questions was reviewed it turned out that one third of them referred to names of individuals or favorites. One sixth of the questions used by one bank were also used by at least one of the other banks. In most cases one third of the questions used by a bank were guessable. 12% of the questions were considered to be automatically attackable, for example in case of dates of birth or postal codes were asked. Rabkin also mentions that none of the banks or websites neither encouraged users to choose or create secure questions nor explained characteristics of a secure question. As a result fallback authentication in online banking is considered to be "surprisingly weak".
Furthermore a survey was conducted to determine users' security awareness related to online banking. In this survey of 46 participants 70% said memorability would be a very important factor when security questions need to be chosen. Only 44% considered security to be an important factor. 7% of participants stated they worry "a lot" about security, 50% worry "some" and 43% "a little". This indicates that security awareness is very low amongst users.
A study by Grossklags and Acquisti [10], evaluating the willingness to protect or sell personal information, supports this theory. It consists of two parts. In the first part of the study participants were asked if they were willing to sell personal information or pay money to protect its release. The second part of the study dealt with the amount of money participants were willing to receive/pay. So the participants were asked to specify the minimum amount of money to release their information respectively the maximum amount of money to protect it. It turned out that the average willingness to sell personal information is much higher [10] than the willingness to pay for protection of this information. In case of weight information 7 of 47 participants were willing to pay for protection. However 41 participants were willing to sell this kind of information. When they were asked if they would sell or protect information referring to the number of sexual partners 34 of 47 participants were willing to sell this information for more money than they were willing to pay to protect it.

Schechter et al. [24] conducted two user studies to analyse the reliability and security of security questions used by four email-providers (AOL, Google, Microsoft and Yahoo!) in the US. 130 participants took part in the first study. Each participant formed a team with a significant-other. The participants were split into four groups whereas the first three groups only consisted of Hotmail users. The members of these groups started by authenticating to their Hotmail account using the corresponding personal question and postal code. Afterwards the participants of all groups were prompted to answer personal questions for the email providers mentioned. These answers had to be guessed by the significant-others. Half of them was given a second chance in guessing the partner's question using online research. Finally the participants attempted to recall their own answers.
It turned out that 42.75% of the Hotmail users could not remember the answer to their security question. 13% of them stated this was due to the fact that they gave a fictious answer at registration. 17% of the answers given by participants in all groups could be guessed by their significant-others. 25% of the answers were considered to be vulnerable for attacks by family members, friends or colleague.
6 months later the participants were asked to recall their answers in a second study. 49 users took part in this study. In this study 20% of the answers could not be recalled.

8 Alternative Question-based Approaches

In their work Jakobsson et al. [13] introduced so-called "dubbed security questions" representing security questions related to personal preferences and characteristics. The mechanism is based on the assumption that preferences and characteristics remain stable over a long period of time. At registration time participants of a study were prompted to answer questions using a three step scale containing the following options: "Really like", "Don't care/Don't know", "Really dislike". Questions and corresponding answers were stored on a server. At authentication participants were asked to answer a subset of the questions used at registration. Thereby he or she was allowed to make mistakes. However if the answers given at authentication vary widely from the answers stored, an attack is likely to be in process. To distinguish between user and attacker a score was computed based on the given answers. Successful authentication of a legitimate user depended on the score being above or below a certain, predefined threshold. However to achieve a tolerable error rate subsets of at least 16 questions were required which results in higher costs for the user. According to the authors further studies are necessary to evaluate the change of preferences over time.

Additional approaches by Haga and Zviran [12, 29] respectively Smith [26] are based on word association. At setup a user creates a list containing a series of cues and answers which is stored by the corresponding service. At authentication the user is asked to answer a subset of cues. To support memorability, cues can be created based on a user-specified topic or context. However to avoid compromise by attackers this topic or context should be kept private.
In a study by Smith [26] this approach was evaluated. Four participants were asked to create a list containing of 20 cues and corresponding answers. Six months later they were asked to recall the list. As they were given the cue list 94% of the answers could be recalled. Without the cues one participant could not recall a single answer. The remaining participants could recall a few answers each. Twelve months later when given the cue list 86% of the recall attempts were successful. It is assumed that in practice recall rate would be higher due to the fact that the cues would be used more frequently. In terms of guessing rate it is said that the success of guessing an answer correctly depends on the context. For example, if the context was known, in one case a close friend could answer four cues, the participant's spouse could answer ten cues. If used in practice it is recommended to test a cue list for its susceptibility to guessing and revise the list accordingly if the guessing rate is too high. However due to lack of experience no further guidelines could be given.

Further mechanisms described by Schechter et al. [24] use authentication codes sent via sms or email to re-authenticate users. Although users don't have to remember certain information like answers to security questions or cues these approaches have several disadvantages. For instance, mobile phones can be lost or stolen. Passwords to access email accounts can be forgotten or compromised. To avoid the latter issue it is recommended to use a secondary email address for receiving authentication codes.

In their work Nosseir et al. [19, 20] use contextual data to authenticate users. Two approaches were introduced. First of all by equipping an office with sensors a user and his or her movement can be detected. Thereby situations can be derived, e.g. arrival or leaving of a user. Based on the idea that each user recurrently moves along a

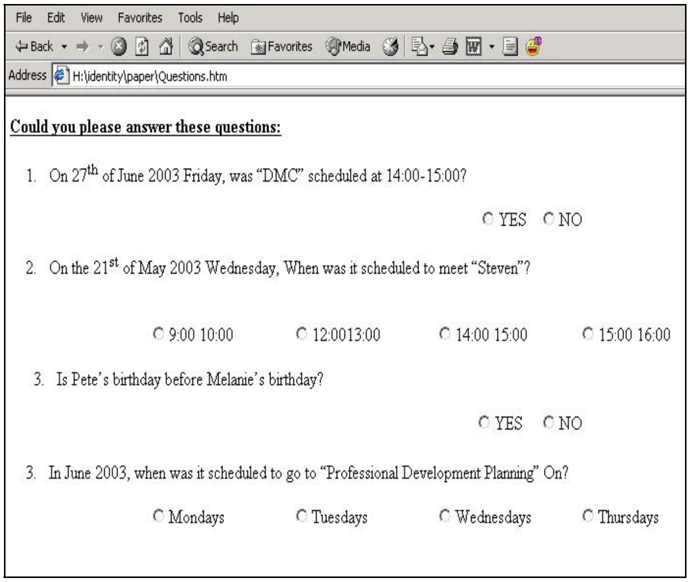


Fig. 2. Questions generated based on calendar events [19]

certain and unique path to reach his or her desk, these situations can be associated to specific users. Due to the fact that these situations are logged, the system can generate questions to authenticate users like "When did you arrive on Wednesday?". In this case to answer the question a range of time intervals could be provided to the user. Although this approach does not require to remember answers to security questions in the common sense or to carry ID keys/cards which can be stolen, certain issues exist that need to be adressed. For example, besides privacy concerns due to the collection of movement data it needs to be further investigated whether definitions for events can be adopted to new users/strangers. Furthermore the amount of questions and options per question respectively the time intervals from which can be chosen are considered to need further research.

The second approach described by Nosseir et al. [19] is based on the use of personal electronic history. By providing a trusted third party data included in this history, e.g. calendar events, authentication questions based on this data can be generated. Figure 2 illustrates generated questions based on calendar data. It is assumed that the amount of data is too large to learn for impostors and hence answers are difficult to guess. As in the first approach by Nosseir et al., users are not required to remember passwords or answers to predefined questions which could be guessed or forgotton. Also, since the approach is knowledge-based physical objects, else used for authentication, cannot be lost or stolen.
In a study data from six calendars was used. It turned out that randomly generated questions could not be answered any better by legitimate users than impostors. However the amount of correctly answered questions and also the recall rate of recent calendar events varied widely from participant to participant due to the fact that the focus was on long-term memory. In a second study data from nine calendars were used. Questions were generated based on recent, repetitive, pleasant (e.g. events being birthdays, parties or concerts) and hence memorable events. Thereby significant results could be achieved. Nevertheless as stated by the authors these events need to be further investigated in a bigger experiment.

9 Conclusion and Outlook

As the title suggests this paper aimed at describing how to design both secure and usable security questions. Therefore different types of questions have been introduced and described with respect to their advantages and disadvantages. Furthermore criteria were presented to support a secure and usable design. Finally the current state of security questions used in practice has been analysed.

Although security questions represent a mechanism widely used, for example by banks [23] or the four biggest email providers in the US [24], as seen in chapter 7, it is very difficult to create questions which are both secure and usable at the same time. It has been shown that there is a trade-off between security and usability. While answers to personal or cognitive questions were easier to remember than conventional passwords they suffered from guessing and observing attacks and could be guessed more easily [12, 22, 30]. Answers to open questions that were assumed to be remembered more easily resulted in persisting recall difficulties [16].

This, on the one hand, is caused by service providers and institutions using less secure questions, for example, by asking for names of individuals or favorites [3, 23]. On the other hand the low security awareness of users, reflected in their willingness to publish personal information in social networks [23] as well as to sell or share these with family or friends [10, 16], hinders protection from those kinds of attacks. However this behaviour might also be caused by a lack of information since service providers and institutions are not providing guidance in how to choose or create secure questions [23]. Thus informing the users will be a crucial aspect in the future.

As mentioned in the previous chapter, besides the question-based authentication systems described in this paper, further alternative approaches exist. On the author's opinion the most promising of these approaches is based on word association. In a study by Smith [26] after 12 months on average 17.2 of 20 answers could be recalled. However the result highly depends on whether a cue list is given or not. To avoid compromise this cue list needs to be kept secret. Furthermore due the fact that only four participants took part in this study additional, extensive investigation is necessary.

Other approaches using context-based authentication [19, 20] suffer from privacy concerns and also need further research, for example in terms of question parameters.

The past has shown that authentication is a crucial aspect of interaction with devices and services. However some forms of authentication, especially question-based authentication, need to adapt to the present. Else alternatives need to be found. Only the future will tell when these alternatives can be found. Though current approaches give occasion for hope.

References

[1] Apple. About Apple ID security questions. `http://support.apple.com/kb/HT5665`. Last Access: May 23, 2014.

[2] Apple. Apple - My Apple ID. `https://appleid.apple.com`. Last Access: May 23, 2014.

[3] J. Bonneau, M. Just, and G. Matthews. What's in a Name? In *Financial Cryptography and Data Security*, pages 98–113. 2010.

[4] T. Bridis. Palin Email Hacker Impersonated Her, Stole Password. `http://www.huffingtonpost.com/2008/09/18/palin-email-hacker-impers_n_127538.html`. Accessed May 4, 2014.

[5] R. Cappelli, D. Maio, A. Lumini, and D. Maltoni. Fingerprint Image Reconstruction from Standard Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1489–1503, 2007.

[6] Facebook. Verwenden deiner Sicherheitsfragen. `https://www.facebook.com/help/www/443380072361197`. Accessed May 23, 2014.

[7] D. Florencio and C. Herley. A Large-scale Study of Web Password Habits. In *Proceedings of the 16th International Conference on World Wide Web*, pages 657–666, 2007.

[8] S. L. Garfinkel. Email-Based Identification and Authentication: An Alternative to PKI? *IEEE Security and Privacy*, 1(6):20–26, 2003.

[9] V. Griffith and M. Jakobsson. Messin' with Texas Deriving Mother's Maiden Names Using Public Records. In *Proceedings of the Third In-*

ternational Conference on Applied Cryptography and Network Security, pages 91–103, 2005.

[10] J. Grossklags and A. Acquisti. When 25 Cents is Too Much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information. In *Proceedings of Sixth Workshop on the Economics of Information Security*, 2007.

[11] P. Gupta, S. Gottipati, J. Jiang, and D. Gao. Your Love is Public Now: Questioning the Use of Personal Information in Authentication. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, pages 49–60, 2013.

[12] W. J. Haga and M. Zviran. Question-and-answer passwords: An empirical evaluation. *Information Systems*, 16(3):335 – 343, 1991.

[13] M. Jakobsson, E. Stolterman, S. Wetzel, and L. Yang. Love and Authentication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 197–200, 2008.

[14] M. Just. Designing and evaluating challenge-question systems. *Security Privacy, IEEE*, 2(5):32–39, 2004.

[15] M. Just. Designing authentication systems with challenge questions. *Security and Usability: Designing Secure Systems That People Can Use*, pages 143–155, 2005.

[16] M. Just and D. Aspinall. Personal Choice and Challenge Questions: A Security and Usability Assessment. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–11, 2009.

[17] M. Just and D. Aspinall. Challenging Challenge Questions: An Experimental Analysis of Authentication Technologies and User Behaviour. *Policy & Internet*, 2(1):99–115, 2010.

[18] A. Moallem. Did You Forget Your Password? In *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, pages 29–39. 2011.

[19] A. Nosseir, R. Connor, and M. Dunlop. Internet authentication based on personal history - A Feasibility Test. In *Proceedings of Customer Focused Mobile Services Workshop at WWW2005*, 2005.

[20] A. Nosseir, R. Connor, C. Revie, and S. Terzis. Question-based Authentication Using Context Data. In *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles*, pages 429–432, 2006.

[21] L. O'Gorman. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, 91(12):2021–2040, 2003.

[22] J. Podd, J. Bunnell, and R. Henderson. Cost-effective computer security: cognitive and associative passwords. In *Proceedings of the Sixth Australian Conference on Computer-Human Interaction*, pages 304–305, 1996.

[23] A. Rabkin. Personal Knowledge Questions for Fallback Authentication: Security Questions in the Era of Facebook. In *Proceedings of the 4th Symposium on Usable Privacy and Security*, pages 13–23, 2008.

[24] S. Schechter, A. J. B. Brush, and S. Egelman. It's No Secret. Measuring the Security and Reliability of Authentication via "Secret" Questions. In *Proceedings of the 30th IEEE Symposium on Security and Privacy*, pages 375–390, 2009.

[25] C. E. Shannon. A Mathematical Theory of Communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001.

[26] S. L. Smith. Authenticating Users by Word Associations. *Computers & Security*, 6(6):464–470, 1987.

[27] H. M. Wood. The Use of Passwords for Controlling Access to Remote Computer Systems and Services. In *Proceedings of the National Computer Conference*, pages 27–33, 1977.

[28] Yahoo! Change your security questions. `https://help.yahoo.com/kb/security-questions-sln2031.html`. Last Access: May 23, 2014.

[29] M. Zviran and W. J. Haga. Cognitive passwords: The key to easy access control. *Computers & Security*, 9(8):723 – 736, 1990.

[30] M. Zviran and W. J. Haga. A Comparison of Password Techniques for Multilevel Authentication Mechanisms. *The Computer Journal*, 36(3):227–237, 1993.

Understanding Users - Implicit Usage Modeling on Smart Mobile Devices

Matthias Lamm

Abstract— Over the last decade, mobile devices with touch screens, commonly known as smartphones, have an increasing market share compared to Personal Computers and laptops. Since these devices are used throughout the whole day in private and business environments, users have an exceptional connection to them. By interacting with them, users generate an enormous amount of information with the installed sensors like gps, illumination or acceleration. With this paper, we want to present current research on how this information can be collected, evaluated, and incorporated in order to support the users. Starting with assumptions which can be drawn for general users, this work further leads to applications for individual user-analysis. Through machine learning and other methods, models of the users can be generated which offer the possibility to predict, compare or assess actions of the user. These models can also be used for biometric authentication in order to evaluate if the current user is the rightful owner or a possible intruder.

Index Terms—Mobile Devices, User Verification, Authentication, App Usage, Machine Learning, Touch screen, Life logging

1 Introduction

Todays smartphones have various opportunities to collect data about their usage for the improvement of the user experience. Diverse sensors like location- or acceleration-sensors or the front and rear camera, and even the touch screen, can deliver information about the specific user. In the first section, we present findings concerning the general behavior of smart-phone-users, and in which way UI/UX designers can use gained information in order to improve the usability of the phone and the apps for users. Usability design on mobile devices is crucial. On the one hand, users want to reach their goals on the fastest way possible. They use their smartphones on-the-go or in crowded places like the subway or for example during short work brakes just to check the latest news or bank account transactions. To reach these goals fast, critical information like mail account information or bank credentials must be stored on the smartphone. On the other hand this critical information has to be well protected against possible threats. Our goal here is to present several possibilities for protecting Smartphones against third parties, and to use easy collectable data in order to improve the usability of certain apps or the smartphone in general. Unlike other approaches, which use dedicated hardware like fingerprint sensors in the iPhone 5s, we want to show possibilities of improvement by data which could be collected by the hardware that is already present in most common smartphones. The second part of this work will focus on how collectable data could improve user specific experiences through implicit authentication methods, the optimization of keyboard input on touchscreens, and automated interface adaptation for smartphones. Finally, psychology related research will be presented. By analyzing the user's behavior with the mobile devices, conclusions to the user's personality can be drawn.

2 General Behavior

In this section we are going to present some investigations concerning the general behavior of users. We show how researchers collect data that helps to analyze the overall behavior of smartphone users and what assumptions can be made for them.

2.1 Which apps are used when

In order to analyze the app usage on mobile devices, the different states and events of an app should be known *(see figure 1)*. While installation and uninstallation normally only appear once, and updating only occurs from time to time, opening and closing are the only events which may be observed regularly. By observing these events, statistics about app usage can be collected.

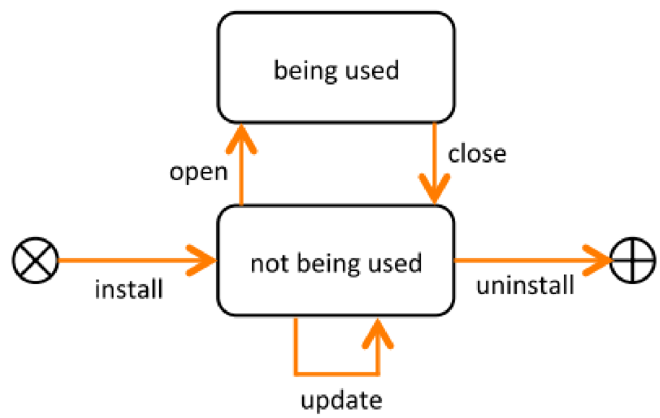


Fig. 1. The five events and two states in the lifecycle of a mobile app: (1) install, (2) open, (3) close, (4) update, (5) uninstall & (I) not being used, (II) beeing used. *(figure from: [3])*

In earlier studies (eg. by Cui and Roto [5]), web usage on mobile devices has been analyzed. Their main findings were that mobile web usage sessions are rather short, compared to stationary web usage via a PC. However when the mobile device is connected to a Wifi, the timeframe for web sessions have been significantly longer, compared to cellular connections. Böhmer et al. [3] started a more detailed analysis of which app categories where most likely to be used at which time *(see figure 2)*. To do so, a data collection tool, AppSensor, has been developed, which has been integrated into an app recommender system app [2].

Their findings show that the overall app-usage per day of the participants is around one hour per day, and the average duration of an application session is 72 seconds. Most apps are started in the afternoon and evening, and the peak is around 6 pm. Contrary, the duration of app usage is the longest in the early morning ours from 3 am to 6 am with five minutes or more. From noon to midnight it's less than a minute. According to Böhmer et al. [3] a possible explanation might be that users explicitly leave certain apps active while they are sleeping and these apps have a standby-mode prevention. When their insights are regarded in more detail, communication apps have the highest probability for a launch during the whole day. In the af-

---

- *Mathias Lamm is studying Human-Computer-Interaction at the University of Munich, Germany, E-mail: M.Lamm@campus.lmu.de*
- *This research paper was written for the Media Informatics Advanced Seminar 'Special Aspects of Usability', 2014*

	12am	1am	2am	3am	4am	5am	6am	7am	8am	9am	10am	11am	12pm	1pm	2pm	3pm	4pm	5pm	6pm	7pm	8pm	9pm	10pm	11pm	% of Total Launches	Users	Apps
Browser	7.9%	7.7%	7.8%	7.6%	7.3%	7.4%	7.0%	7.9%	8.1%	8.0%	7.7%	7.3%	7.0%	6.9%	6.8%	6.4%	6.6%	6.6%	6.4%	6.6%	7.0%	7.4%	7.5%	7.4%	6.83%	2,398	9
Comics	4.5%	5.2%	5.4%	5.8%	5.8%	5.6%	5.5%	5.2%	5.4%	5.1%	4.7%	4.3%	4.3%	4.2%	4.2%	4.3%	4.4%	4.0%	4.4%	4.2%	4.1%	4.1%	4.1%	4.4%	4.31%	2,151	1,810
Communication	44.9%	41.1%	38.3%	35.4%	31.6%	31.8%	32.7%	34.7%	39.4%	44.8%	49.0%	52.6%	54.8%	55.2%	55.2%	56.1%	55.7%	56.8%	57.1%	56.1%	54.8%	53.3%	52.0%	49.0%	49.50%	2,769	550
Entertainment	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.02%	126	43
Finance	0.2%	0.3%	0.3%	0.2%	0.1%	0.1%	0.1%	0.2%	0.3%	0.3%	0.4%	0.5%	0.3%	0.3%	0.4%	0.3%	0.3%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.25%	604	164
Games	3.2%	3.0%	3.0%	2.7%	2.5%	2.3%	2.2%	1.7%	1.9%	1.9%	2.0%	2.1%	2.2%	2.2%	2.2%	2.3%	2.3%	2.2%	2.2%	2.4%	2.7%	3.0%	3.0%	3.2%	2.30%	1,716	1,702
Health	0.3%	0.4%	0.4%	0.4%	0.6%	0.6%	0.7%	0.6%	0.4%	0.3%	0.3%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.3%	0.2%	0.3%	0.2%	0.3%	0.26%	540	227
Libraries & Demo	0.4%	0.5%	0.6%	0.7%	0.9%	0.8%	0.7%	0.6%	0.5%	0.4%	0.3%	0.3%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%	0.30%	1,267	117
Lifestyle	0.8%	0.9%	1.0%	1.4%	1.3%	1.5%	1.4%	1.4%	1.1%	0.9%	0.6%	0.6%	0.5%	0.5%	0.5%	0.5%	0.6%	0.5%	0.3%	0.4%	0.4%	0.5%	0.5%	0.5%	0.60%	2,132	451
Multimedia	2.1%	2.1%	2.4%	2.4%	2.7%	2.4%	1.8%	1.8%	1.9%	1.7%	1.8%	2.0%	2.0%	2.0%	2.2%	2.1%	2.2%	2.4%	2.3%	2.3%	2.2%	2.1%	1.9%	2.0%	2.03%	1,713	76
News	2.6%	2.5%	2.6%	2.5%	2.5%	2.7%	3.3%	3.7%	4.1%	3.6%	3.0%	2.6%	2.5%	2.7%	2.5%	2.4%	2.2%	2.1%	2.3%	2.2%	2.3%	2.2%	2.3%	2.3%	2.46%	1,777	440
Productivity	3.6%	5.0%	5.0%	5.8%	6.3%	6.5%	6.0%	5.4%	4.8%	5.1%	4.9%	4.3%	4.2%	4.0%	4.0%	3.7%	3.4%	3.4%	3.0%	3.1%	3.1%	3.0%	2.9%	3.2%	3.76%	2,190	648
Reference	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.6%	0.6%	0.7%	0.5%	0.5%	0.5%	0.4%	0.4%	0.4%	0.4%	0.3%	0.4%	0.4%	0.4%	0.5%	0.5%	0.5%	0.6%	0.47%	903	346
Settings	1.3%	1.6%	1.5%	1.3%	1.6%	1.2%	1.2%	1.1%	1.3%	1.4%	1.4%	1.4%	1.2%	1.3%	1.2%	1.2%	1.3%	1.1%	1.1%	1.2%	1.2%	1.3%	1.3%	1.4%	1.23%	2,178	1
Shopping	3.9%	4.5%	3.7%	3.4%	3.2%	3.2%	3.1%	3.0%	3.1%	3.3%	3.2%	3.2%	3.2%	2.8%	2.9%	2.9%	2.7%	2.7%	2.7%	2.7%	2.8%	3.1%	3.6%	3.5%	2.96%	2,556	198
Social	5.7%	5.0%	4.9%	4.3%	4.2%	4.0%	4.4%	5.1%	5.3%	5.4%	5.2%	5.0%	4.7%	4.8%	4.9%	4.5%	4.5%	4.6%	4.6%	4.9%	5.2%	5.4%	5.8%	5.7%	4.77%	1,902	342
Sports	0.5%	0.3%	0.3%	0.2%	0.3%	0.3%	0.2%	0.3%	0.3%	0.3%	0.3%	0.4%	0.4%	0.6%	0.7%	0.8%	0.9%	0.8%	0.6%	0.6%	0.7%	0.8%	0.7%	0.7%	0.56%	571	215
Themes	0.2%	0.1%	0.2%	0.3%	0.4%	0.4%	0.4%	0.2%	0.2%	0.2%	0.1%	0.1%	0.1%	0.2%	0.1%	0.1%	0.1%	0.2%	0.1%	0.1%	0.2%	0.1%	0.1%	0.1%	0.14%	249	231
Tools	10.9%	12.2%	14.6%	17.6%	20.3%	21.5%	21.4%	18.6%	14.7%	10.4%	8.4%	6.8%	6.1%	5.9%	5.9%	5.9%	6.0%	6.1%	5.8%	6.0%	6.3%	6.8%	7.4%	9.1%	7.89%	2,512	1,688
Travel	1.4%	1.6%	2.1%	2.2%	2.4%	2.6%	2.2%	1.9%	2.0%	2.1%	2.0%	1.8%	1.9%	1.9%	1.9%	1.8%	2.0%	1.9%	2.2%	2.2%	1.9%	1.7%	1.6%	1.4%	1.86%	1,752	407
Unknown	4.7%	5.3%	5.1%	5.0%	5.3%	4.4%	5.0%	5.9%	4.6%	4.4%	4.1%	3.8%	3.5%	3.8%	3.7%	3.7%	4.0%	3.6%	3.7%	3.7%	3.7%	3.9%	4.1%	4.5%	3.88%	2,284	1,796
Total Launches per Hour	103,604	77,053	53,633	40,332	33,438	30,949	38,161	56,895	83,488	109,550	127,069	142,642	158,876	168,082	169,018	172,935	173,963	179,801	184,012	176,050	163,080	153,835	141,303	123,639			

Fig. 2. Representation of the relative app launch frequency. Each app is assigned to one of the categorys shown in the rows. The columns represent one hour time intervals through the day. Each cell shows the relative usage of apps of the respective category in the time interval. Through normalization of the rows the colour-coding is realized. Dark green represents the highest values of each row while white stands for the lowest. So one can for example easily see that communication apps are most likely launched in the afternoon and less often in the early morning. *(figure from: [3])*

ternoon and evening hours the probability is even above 50% of all launched apps. This shows that mobile devices are, despite their wide spread usage opportunities, still mainly communication devices. News apps in contrast, have their launch peak in the morning hours. Insights like these are also used by commercial news networks. Yahoo for instance just launched "News Digest", a curated and algorithm driven news app, which presents eight to twelve news twice a day at 8 am and 6 pm via notification[1]. This is an aid for users, who now don't need to search for their news themselves, avoiding distractions or work interruptions. Another aspect are "application chains": They are defined as an sequence off apps which are used on the device while the standby-mode is not active for longer than 30 seconds [3]. Most sessions (68.2%) only contain a single app. In multi-app sessions users reached a maximum of 14 unique apps but tend to reuse previously used apps. Nearly every second multi-app session was initialized with an app from the communication category. This again emphasizes the role of mobile devices as communication devices. Besides time, location is a further covariable for the usage behavior with apps, according to empirical evidence. As soon as users are moving with speeds higher than 25 km/h, they most likely use Multimedia-apps. Presumably they use their mobile device to listen to music in the car or the bus. When users are located at airports, the probability of them using a browser app is nearly three times higher. Böhmer et al. [3] concludes that it's possible that this is a sign for the lack of native apps, which provide the desired information like flight status, forcing users to use conventional websites.

Since the study is from 2011, some findings probably must be relativized. The range of apps in the appstores of Apple, Google and Microsoft is growing every day. Airlines for example offer their own apps which provide the users with a better experience than the mobile website of the airline. Parallels can also be drawn for other fields, like for example web mail providers or news sites. It can be assumed that the release of dedicated smartphone apps raises the reach on mobile devices. If a company determines their website is often used with mobile devices, one can advise that the development of an app might be a good idea. After users install such a new app on their device the next step is the app launch.

[1] http://www.theverge.com/apps/2014/5/8/5692760/yahoo-news-digest-arrives-on-android-and-launches-an-international (Last accessed: 2014-05-31)

2.2 How Users Launch Apps

Due to the mentioned increasing number of apps that users have installed on the phones, it's crucial for users to organize their screen in order to launch the desired apps as fast as possible. Regarding the android operating system, users have several possibilities to start apps:

- Launcher: Home-screen + Drawer + Folders + Widgets + Dock
- Notification Bar
- Within other apps
- Lock-screen

Fig. 3. Screenshot of android homescreens with the different areas from which apps can be started: (1) Dock, (2) App Drawer, (3) Panel, (4) Widget & (5) Notification Bar. The left screenshot shows the original homescreen and the right one the corresponding reconstruction for the userstudy by Hang et al. . *(figure from: [11])*

To analyze the launch and organize behavior, Hang et al. developed a custom android launcher, which made it possible to collect this information. Gaining knowledge about this launch behavior *"is important* *to identify areas for improvements and to understand their implications for the design of app launchers."*[11] In order to provide real life data, the launcher was installed on the user's own devices, and their homescreens were reconstructed as far as possible *(see figure 3)*. During the study, data of 22 participants was collected for four weeks. Investigating the collected information, revealed that most **app starts** were performed from the launcher and from within other apps through intents. The notification bar has only been used by few users. Only one participant frequently used the lock screen to launch apps. Additional **navigation time** has been logged. When a navigation type (for instance the app drawer) was opened, timekeeping got initialized and stopped when an app was launched. The measured average navigation intervals reached from two to five seconds [11]. Obviously, dock-launches offer the fastest average navigation time, followed by the homescreen, folders and at last the app drawer. Remarkably, horizontal app drawers performed 20% slower than vertical ones. Regarding the **arrangement**, users surprisingly rarely use the dock for their most used apps, despite this being the fastest possibility. Some users even used the app drawer to launch some of their most frequently used apps. This leads to the assumption that there is much space for possible improvements. *"Due to the individual traits, the design of app launchers should not follow a 'one design fits it all' approach, but should take users personal preferences into account (i.e. using existing launching concepts and complement them with adaptive solutions)."*[11]

Most of the possible adjustments and settings here lie in the hands of the users himself and app developers do not have many opportunities to influence them. However, on Android the possibility to install alternative launchers with automated settings does exist. These will be addressed later. Now we are going to look at how users interact with the touchscreen of their smartphones when an app is running.

2.3 Touchscreen Usage Analysis

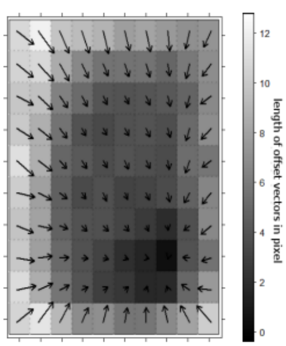


Fig. 4. The touchscreen is divided in a 10x10 grid. The arrows represent the relative offset between targets, which are in the respective cell and the actual tapping point. The brighter each cell, the longer the offset vector. At the far left targets show the highest offset. *(figure from: [12])*

When regarding the usability of touchscreens, you have to consider how the user is interacting with it. Depending on the situation, there are three possible positions how users operate with mobile devices:

- one handed
- two handed - one hand touching
- two handed - both hands touching

In each position you have to consider that different areas of the screen have different values concerning the reachability. Additionally, touchscreen inputs can be influenced by the "fat finger problem" where the desired target gets occluded by the finger, and a precise input is impeded. Therefore targets like buttons must have an appropriate size to reduce error rates.

Henze et al. [12] developed a game for mobile devices, where the user has to tap appearing targets in the fastest way possible. With increasing progress the levels of the game become more difficult, up to a nearly unsolvable level. They published the game in the android market to reach a wide audience. The performance of the users got recorded and transmitted to the developers. The collected data shows, that depending on the location on the screen, different areas show an offset between a target and the actual tapping point *(see figure 4)*. Besides the position of the targets, the target size has also been considered. In a second version of the game, a location dependent shift-function has been included. Results show a significantly reduced error rate. Through this, a systematically skewed input on touchscreen mobile devices could be proven.

Generally we can assume that areas near the edges of a touchscreen are difficult to reach with a high accuracy. "Sweet-spots" with high accuracy are dependent to the holding position and the screen size. Since Android devices have various touchscreen-sizes and resolutions and therefore positioning on the screen has to be relative, android developers have a disadvantage compared to iOS developers where size and resolution is fixed.

3 Individual Behaviors

While the previously shown research is looking for globally valid findings, we now seek to show works that regard the individual users and their behavior.

3.1 Biometric User Authentication Approaches

With the increasing spread of smartphones in today's society, security problems concerning your smartphone gains more and more importance. The most used protection for phones are the four digit PIN, and the android unlock pattern *(see figure 5)*, which has been introduced by Google in 2010. For the PIN-system, *"the number of candidate passwords is limited to only 10,000 (from 0000 to 9999)"*[13] possibilities. These threats will be regarded now in more detail.

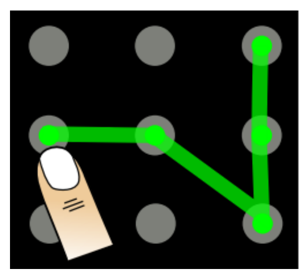


Fig. 5. The Android unlock pattern: Dots on a 3x3 grid get connected by drawing a line on the screen with the finger. The smartphone only gets unlocked if the right dots get connected in the right order. *(figure from: [18])*

Bruteforce: Since there are only 10.000 possibilities for the four character pin chances to unlock a phone just by guessing are not that bad. If there's personal knowledge about the owner, for example the birth date, the risk of an unwanted unlock is even higher.

Smudge attacks: Aviv et al. [1] took a look at how the widely held android unlock pattern could be easily attacked just by photographing the Touchscreen and adjusting brightness and contrast of the image. Due to oily residues, the used pattern can be reconstructed even when "noise" by normal usage is generated, or when tracks are blurred by cloth contact. To prevent this von Zezschwitz et al. [18] tested alternative authentication modes.

Shoulder-surfing: Since phones are often used in public, shoulder-surfing is another possible threat for the security of the smart phone. *"Shoulder-surfing refers to someone watching over the user's shoulder as the user enters a password, thereby capturing the password. While alphanumeric passwords systems are vulnerable to shoulder-surfing if the attacker can see the keyboard, graphical password systems may be more vulnerable in certain settings."*[19]

Implicit Authentication is a method to authenticate users by actions they would perform anyways [14]. Jakobsson et al. proposed that implicit authentication could be used as

Fig. 6. Four different unlock screens used by de Luca et al. in their user study. Horizontal swipe, vertical swipe, vertical swipe with two fingers and diagonal swipe unlock. *(figure from: [6])*

- an extra security level, which would be used alongside common password systems, or
- as the only method of authentication, because users could be annoyed of using passwords, or
- it could just be used during critical actions, like for credit card payment transactions.

They categorized data into three groups: **Device data**, which is given by the used device like GPS coordinates, accelerometer values, WiFi, Bluetooth an USB connections, but also activity of other applications, which are active on the Device. Additionally, other biometric measurements like keystroke dynamics, which will be regarded later, could also be used. **Carrier data** can contain information about cellular location coordinates, call logs or even voice samples that can be matched with the current user. **Cloud data** is provided by hosters and give other useful information, like for example calendar entries.

By combining this information, a model of the user can be generated, with which actions of the users get evaluated. If it's a typical user action, like calling the parents, a score gets increased. Otherwise, if it's an untypical action, like a wrong authentication, this score will be decreased. Additionally, the score gets degraded over time. As long as this score is above a certain threshold, the device could be used without any further required actions. If the score falls below this threshold, the user has to re-authenticate with his password, which further increases the score.

Another approach for using biometric authentication has been used by De Luca et al. [6]. Their goal was to determine how users behave during the login process. The basic idea was to exploit touch screen data of common smartphones (without adding additional hardware) in order to identify users based on how they perform during the login. Two studies have been carried out.

The first one regarded different unlock screens *(see figure 6)*. The permutations of lockscreens contained of a horizontal, a vertical with one and two fingers, and a diagonal gesture. The participants were assigned to the different types. During the study, all information of the touchscreen has been collected: the pressure, size, x- and y-coordinates, as well as the time. Each user had to train the system how they unlock their phone. This model was later compared to the other users via dynamic time warping (DTW). With this algorithm, which originates from speech recognition [16], the costs for matching two different sets is calculated. If these costs are below a predetermined threshold, access is granted. Whereas true positives have been above 90% in most cases, the accuracy is around 50% due to high false positive rates (excluding the two finger vertical condition, which is significantly lower).

The second study considered the commonly known password patterns. In contrast to the first study, more complex data could be collected because the pattern normally consist of more complex gestures. After a field study, in which users had to train their model with a predetermined pattern over 21 days, they had to simulate attacks on the patterns of other users in a lab session. Due to these alternations, the accuracy could be improved to 77%. *"The results of the second study support our claim that password patterns create data that is distinct enough to distinguish between different users. Overall, it can be stated that using touch screen data to identify users works to a certain degree."*[6]

Another attempt for biometric authentication are **Keystroke dynamics**. To improve data security besides of the knowledge (like password) or possession (like ID-cards), which can be easily compromised or stolen another level of security is introduced: **Biometrics** are features that define a specific person. These features can be physiological like for example body- or face- shapes (used for example in face-recognition), but also behavioral like typing rhythm or speech. An advantage against other methods is that these features can not be lost and are mostly difficult to imitate for potential intruders. This makes Biometrics a good identifier for authentication processes.[15]

With keystroke dynamics, the user's typing behavior on the keyboard gets analyzed. Studies have shown that the keystroke rhythm is a good indication for someones identity. Monrose and Rubin [15] ran their studies on different workstations at NYU and Bell communications. Their system used euclidean distances to distinguish different datasets, and the used classifier reached a correct identification rate of 87.18%. Besides their goal, to identify certain users, they made the suggestion that keystroke dynamics could also be used at workplaces where high awareness is crucial, like for air traffic controllers because drowsiness or fatigue could be recognized and third parties could be alerted.

Keystroke Analytics on Mobile Devices

While Monrose and Rubin[15] just took desktop systems into consideration, Frank et al., Hwang et al. and Zeng et al. [10, 13, 22] transferred the approach to cell phones and smart-phones *(see figure 7)*

Touchscreens of modern smartphones and the modern mobile OSs give various additional information, which can be used for feature generation. Zheng et al. [22] added acceleration, pressure, size and time of touch-events during logins to generate unique user patterns. Analysis of their data showed that combination of all these features outperforms the individual features.

Hwang et al. [13] developed a framework for mobile devices, which also used the password authentication for an additional biometric verification. After an enrollment of valid users, a classifier is build. The access to the device can be denied to unknown users who gained ac-

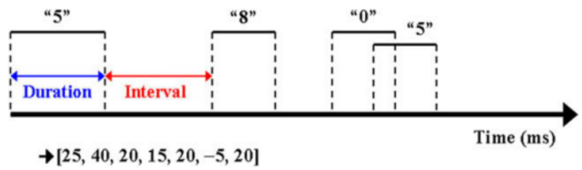


Fig. 7. The string '5805' is typed by a user. The duration (blue) is the length of the 'key down' event on the keyboard. The interval (red) is the time between two 'key down' events. At the bottom the resulting timing vector is shown: It consist of tupels of durations and intervals.The first number is the duration of a 'key-down' event and the second one the interval to the next 'key-down' event. The negative interval shows overlapping 'key-down' events where both keys are pressed at the same time. *(figure from: [13])*

cess to the correct password but use the wrong input rhythm *(see figure 8)*. Again a simple distance measurement is used to match the authentication try with the trained model. In their study, they used two conditions: "Natural Rhythm without Cue" and "Artificial Rhythms with Cues". Cue meaning a metronome-like indication, to support the users to use the correct rhythm. Their results show that error rates for valid users could be decreased in most cases.

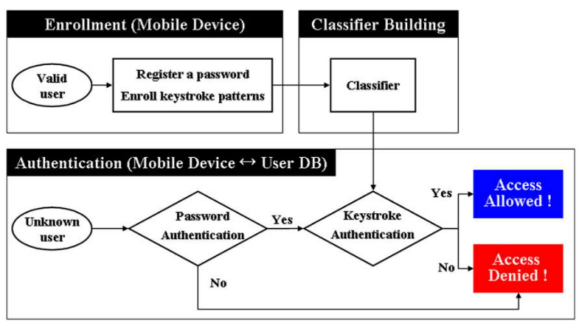


Fig. 8. Diagram of the framework used by Hwang et al. showing the three steps: enrollment - a new user gets registered, classifier building - the modell for the respective user gets build, user authentication - if a user enters a password the input gets matched with the classifier before granting access. *(figure from: [13])*

Security on smartphones gains more and more importance and the previously shown approaches have shown that security can be improved by userdata collection and evaluation. However Apple and Google are tending towards building new hardware to improve security. While Apple is using Touch ID[2], Google presented a technique at their developer conference Google I/O, where the physical presence of other hardware, like for instance a smart-watch, automatically unlocks the smartphone[3].

3.2 Typing Optimization

A model of specific users may not only be used to identify him or her, but also to support them when it comes to text input. Earlier we showed that communication is still the most used task for phones. Besides speech, text input through keyboards, is commonly used.

Systematic touch offset, like shown earlier is also a user individual problem, when it comes to every form of touch input like on soft-keyboards, which are used on most of todays mobile devices. Besides the problem that keyboards on mobile devices are multiple times smaller than desktop or laptop keyboards, they lack haptic feedback, which is sometimes simulated via the vibration motor of the devices. However since they are software-driven user specific adaptations can be performed to improve input accuracy. Findlater et al. [9] made a study on how adaptive keyboards perform against static keyboards, which only use the raw input data. Their adaptive keyboards pursued two conditions: While one adapts the classification model, only in the invisible underlying model, the other one also adapts the visible keyboard *(see figure 9)*.[9] For the training a J48 classifier was used. This is an open source implementation of C4.5, a decision tree algorithm. During the training, the underlying keyboard model got updated after every entered sentence. Contrary to the expectations, the keyboard with visual adaptation provided no improvement, while the not visually adapting keyboard improved the typing speed up to 15.2% in the third training session of the study. Regarding error rates, however no perceivable differences could be found. Back-of-Device keyboards can also benefit of such a user specific adaptation. Buschek et al.[4] presented a Clustering-based approach to improve the performance of users during input. In contrast to the previous portrayed studies of Findlander et al.[9], they also used additional language models to predict inputs. They combined both keyboard- and hand-models for a hierarchical clustering method. Hereby a re-estimation for the locations of keys can be performed dynamically while typing.[4]

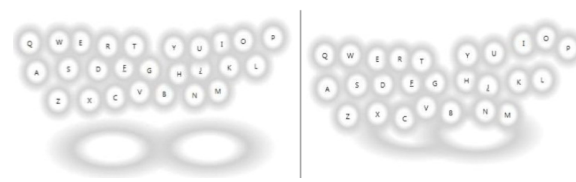


Fig. 9. Visually adapted layout of an adaptive keyboard for two different users after they had several training sessions. *(figure from: [9])*

For improvements during the typing process it is essential to build a good text-model for the specific user. To speed up the learning process of the model and adapt the "language style" of the user, many keyboards like swiftkey[4] want permissions to read previously typed texts like emails or social network statuses. However improvements for the model are also performed dynamically.

3.3 User Interface Adaptation

Since users, like shown before, often tend to not use their mobile devices in the most effective way and waste time by searching for their desired app, it's a viable field of research when it comes to usage prediction. If users get presented the most probable app automatically, they don't have to waste time searching through their increasing number of apps. Such an approach is followed by Shin et al. [17] and similar at Xu et al. [20].

Again, a logging app was designed which collected User-related- (GPS, cellular network location, 3D accelerometer, personal schedule, calls and SMSs), Environment-related- (Illumination, carrier, Wifi, Bluetooth, screen status, battery status, setting status, and device status), and App-related- (running apps, active app, app status) -information. In their study, they collected sample-data with context information for launched apps and formed an inference model with naive Bayes classifiers (offers good tradeoff between accuracy and computation time). With this model, they could calculate the probability for an app start in a given context.[17] Often other methods like Support Vector Machines (SVM) offer better results. Since these have enormously higher requirements, concerning processing time and memory, their usage is not advisable on mobile devices.

The developed model was later used to build a launcher app which dynamically reorganizes the app icons on the first home screen. The icons were organized from top left to bottom right in decreasing order and the app with the highest gain in probability is color-highlighted. Their results showed that the participants frequently used the new homescreen. The search time for apps which were normally started by the users via the app drawer decreased understandably. Apps which

[2]http://support.apple.com/kb/HT5883 (Last accessed: 2014-06-30)
[3]http://www.theverge.com/2014/6/25/5842050/the-most-important-things-google-announced-at-i-o (Last accessed: 2014-06-30)
[4]http://swiftkey.com/ (Last accessed: 2014-06-30)

	Datatypes	Datacollection	Study Type
Which Apps	App starts, Location, Time etc.	AppSensor	Field
Launch Behavior	App starts, Touch coordinates, Package events, Time etc.	Launcher	Field
Touchscreen Usage	Touch coordinates, Target size an position, Resolution etc.	Touchgame	Field
Biometric Authentification	Touch coordinates, Acceleration, Pressure, Time etc.	e.g. Lockscreen apps	Lab and Field
Typing Optimization	Touch coordinates, Keyboard usage, Dictionaries etc.	Keyboard apps (prototypes)	Lab
User Interface Adaptation	Location, Time, Acceleration, Wi-Fi etc.	Adaptive Launcher	Field

Table 1. Presented research fields and collected data

were previously on the static homescreen, now required slightly more time, because users had to search through the reordered homescreen and can not rely on their memory.

Similar attempts are today already used in productive systems like Google Now[5] on android devices. Current location, information from Googles social network Google+ and search history is used to present "cards" with information the user might need at the moment. Aviate[6] an android launcher currently in beta phase, adapts the homescreen accordingly to preset locations like work and home, time dependent or for instance if the user plugs in headphones. In each situation groups of apps, which can be linked to this situation, are shown. These groups themselves have a fixed order, which could be adjusted by the user to avoid confusion.

4 Conclusion and Discussion

Regarding the presented researches, one can see that there are various fields which are investigated today. Table 1 offers an overview for the addressed topics. Most approaches use similar data types which are mostly gathered from the smartphones sensors but also directly from apps like the dictionaries from used words. A problem is that researchers always have to implement their own data collection. However, the Android OS for example offers the possibility to use various intends which can be used system wide, but the storage for such information is application specific. Since the same data types are used often for various investigations, it would be a great opportunity if the OS manufacturers like Google and Apple grant access to a usage information storage for developers and researchers. By this, redundantly stored information could be avoided and developers can still use it. Hereby it would probably be easier to gather information with productive systems instead of being dependent on labstudies often times. Another approach would be the development of a framework that collects and holds the information, and can be accessed by other apps.

Nevertheless this data is also critical like user credentials and must be well protected. Especially in Europe and Germany, users and the media often see it as being problematic when it comes to user data collection. To minimize the user's fear of abuse, data protection and anonymization must be granted.

5 Outlook

These previously shown systems are mostly used to improve the user experience for the currently used device. However data which could be collected by mobile devices can also be used to generate more abstract personality models of the users themselves. Yuruten [21] tried to derive predictors of life satisfaction via these data. To do so, communication information (how many calls or SMS conversations are initialized/received by the user), proximity information (how many other devices with active bluetooth are around the user), and location (home/work/elsewhere) are used. This information got compared with conventional surveys for social life satisfaction *(see figure 10)*. Their *"results show that the method is useful for estimating user activities, and for identifying meaningful relations between activities and satisfaction. More specifically, the analysis shows that work, leisure and sleep activities, and regularities in the daily activities have both direct and indirect inuences over the reported levels of satisfaction."*[21]

[5]http://www.google.com/landing/now/ (Last accessed: 2014-06-23)

[6]http://www.theverge.com/2013/10/15/4839162/aviate-android-home-screen-app-google-now (Last accessed: 2014-05-31)

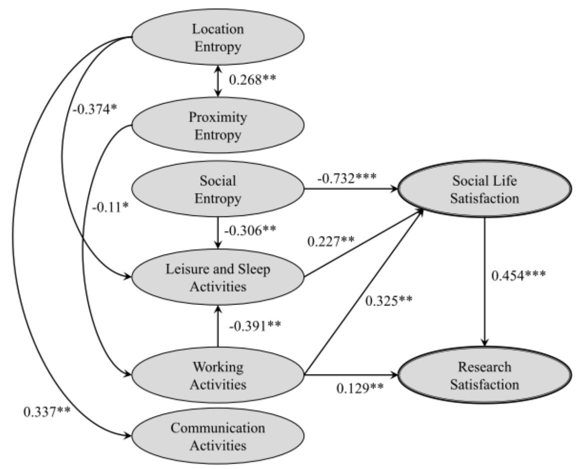


Fig. 10. Modell generated by Yuruten et al. which shows the found relationships. For example a high social entropy has a negative effect on the social life satisfaction. The *, ** and *** stand for the significance levels p <0.1, p <0.05 and p <0.001 of the respective path. *(figure from: [21])*

De Montjoye and De Oliveira [7, 8] have even made an approach, to predict peoples personality. De Montjoye[7] build a model which *"predicted whether phone users were low, average, or high in the big 5: neuroticism, extraversion, conscientiousness, agreeableness, and openness."*[7]

Currently, we are developing an android app, which will log the users interactions with the device. Our participants will perform conventional intelligence and personality tests. Afterwards the app will be installed on the user's own devices, which will send the collected data to our servers. Our goal will be to confirm [7] findings and develop a machine learning approach, which makes it possible to predict a certain user's personality and probably even his or her intelligence from their interaction with the smartphone.

References

[1] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith. Smudge attacks on smartphone touch screens. In *Proceedings of the 4th USENIX conference on Offensive technologies*, pages 1–7. USENIX Association, 2010.

[2] M. Böhmer, G. Bauer, and A. Krüger. Exploring the design space of context-aware recommender systems that suggest mobile applications. In *2nd Workshop on Context-Aware Recommender Systems*, 2010.

[3] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In *Proceedings of the 13th international conference on Human computer interaction with mobile devices and services*, pages 47–56. ACM, 2011.

[4] D. Buschek, O. Schoenleben, and A. Oulasvirta. Improving accuracy in back-of-device multitouch typing: A clustering-based approach to keyboard updating. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, pages 57–66, New York, NY, USA, 2014. ACM.

[5] Y. Cui and V. Roto. How people use the web on mobile devices. In *Proceedings of the 17th international conference on World Wide Web*, pages 905–914. ACM, 2008.

[6] A. De Luca, A. Hang, F. Brudy, C. Lindner, and H. Hussmann. Touch me once and i know it's you!: implicit authentication based on touch screen patterns. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 987–996. ACM, 2012.

[7] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland. Predicting personality using novel mobile phone-based metrics. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 48–55. Springer, 2013.

[8] R. de Oliveira, A. Karatzoglou, P. Concejero Cerezo, A. Armenta Lopez de Vicuña, and N. Oliver. Towards a psychographic user model from mobile phone usage. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 2191–2196. ACM, 2011.

[9] L. Findlater and J. Wobbrock. Personalized input: improving ten-finger touchscreen typing through automatic adaptation. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 815–824. ACM, 2012.

[10] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *Information Forensics and Security, IEEE Transactions on*, 8(1):136–148, 2013.

[11] A. Hang, A. De Luca, J. Hartmann, and H. Hussmann. Oh app, where art thou?: on app launching habits of smartphone users. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pages 392–395. ACM, 2013.

[12] N. Henze, E. Rukzio, and S. Boll. 100,000,000 taps: analysis and improvement of touch performance in the large. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 133–142. ACM, 2011.

[13] S.-s. Hwang, S. Cho, and S. Park. Keystroke dynamics-based authentication for mobile devices. *Computers & Security*, 28(1):85–93, 2009.

[14] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit authentication for mobile devices. In *Proceedings of the 4th USENIX conference on Hot topics in security*, pages 9–9. USENIX Association, 2009.

[15] F. Monrose and A. D. Rubin. Keystroke dynamics as a biometric for authentication. *Future Generation computer systems*, 16(4):351–359, 2000.

[16] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[17] C. Shin, J.-H. Hong, and A. K. Dey. Understanding and prediction of mobile application usage for smart phones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 173–182. ACM, 2012.

[18] E. Von Zezschwitz, A. Koslow, A. De Luca, and H. Hussmann. Making graphic-based authentication secure against smudge attacks. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 277–286. ACM, 2013.

[19] S. Wiedenbeck, J. Waters, L. Sobrado, and J.-C. Birget. Design and evaluation of a shoulder-surfing resistant graphical password scheme. In *Proceedings of the working conference on Advanced visual interfaces*, pages 177–184. ACM, 2006.

[20] Y. Xu, M. Lin, H. Lu, G. Cardone, N. Lane, Z. Chen, A. Campbell, and T. Choudhury. Preference, context and communities: a multi-faceted approach to predicting smartphone app usage patterns. In *Proceedings of the 17th annual international symposium on International symposium on wearable computers*, pages 69–76. ACM, 2013.

[21] O. Yürüten, J. Zhang, and P. H. Pu. Predictors of life satisfaction based on daily activities from mobile sensor data. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 497–500. ACM, 2014.

[22] N. Zheng, K. Bai, H. Huang, and H. Wang. You are how you touch: User verification on smartphones via tapping behaviors. Technical report, Tech. Rep. WM-CS-2012-06, 2012.