Cross-Modal Natural User Interfaces for Mobile Devices: From Sensory Data to Augmented Cognition

ADVANCED SEMINAR
by

Thomas Larasser

29.04.1992

Gabelsbergerstraße 70
80333 München
Tel.: 0173 1877728

Neuroscientific System Theory
Technische Universität München

Prof. Dr. Jörg Conradt

Advisor: M.Sc. Cristian Axenie
Submission: 18.01.2016

In your final hardback copy, replace this page with the signed exercise sheet.

Abstract

Over the past years an increasing performance in sensing and computational power of the pervasive and ubiquitous mobile devices resulted in the development of more and more complex and multifarious interfaces which are demanding not only in their implementation but also for the users. Challenges are the integration of multimodal interaction while considering cognitive neuroscientific research in order to create interfaces which provide powerful and natural interaction experience. This work reviews the key aspects for the development of multimodal natural user interfaces on mobile devices. First a model of the cerebral sensory integration and some essential brain mechanisms are presented extracting the role and restrictions of crossmodality in the human brain. Then an overview over the core design principles especially regarding the role of context and the implementation of multimodal fusion is given. Finally it will be discussed how the limitations of mobile devices influence the implementation and the document concludes stressing the main remaining challenges and outlining the future of Natural User Interfaces.

Zusammenfassung

In den vergangenen Jahren führte die Steigerung von Mess- und Rechenleistung vor allem in den allgegenwärtigen mobilen Geräten zu immer komplexeren und vielseitigeren Benutzeroberflächen. Diese stellen hohe Anforderungen, sowohl an die Implementierung, als auch die Benutzer selbst. Eine besondere Herausforderung ist die Integration mulitmodaler Interaktion unter Berücksichtigung kognitiver neurowissenschaftlicher Forschungsergebnisse.

Diese Arbeit überblickt die Schlüsselaspekte für die Entwicklung natürlicher multimodaler Benutzeroberflächen. Zuerst wird ein Modell der cerebralen Integration der Sinne und einige Hirnmechanismen vorgestellt mit der Absicht die Cross-Modalität des menschlichen Gehirns und dessen Einschränkungen aufzuzeigen. Darauf folgt ein Überblick der Kernprinzipien des Designs mit besonderem Bezug zur Rolle von Kontext und der Durchführung von Sensordatenfusion. Außerdem wird noch diskutiert inwiefern die Grenzen mobiler Geräte die Implementierung beeinflussen mit einer abschließenden Betrachtung zukümtiger Herausforderungen von natürlichen Benutzeroberflächen.

Contents

Chapter 1

Introduction

Nowadays most everyday activities contain interaction with computational devices. More frequently in certain areas as the working environment but with the advent of smart homes even traditional simple tasks as switching on lights or controlling the heating of a room require the handling of more complex tools than switches.

A sophisticated interface design is needed which not only enables controlling but is also supportive to the end user in order to facilitate the interaction between human and computer (HCI). Due to technological restrictions this was not the main focus in historical user interfaces where the designer was glad to reach some level of functionality. Therefore the first one the Command Line Interface (CLI) required high expertise and knowledge of the system in use. Interaction itself was constrained using only text input and the user had to recall the instructions necessary to direct the program. With the improvement of computer systems in the fields of memory, processing power and visualization technology the Graphical User Interface (GUI) emerged. The underlying WIMP (Windows, Icons, Menus, Pointers) - concept was essential for its success in the past years. Especially non expert users were now able to stir complex computational systems in an indirect and exploratory way using graphical and more natural representations.

Further technology improvement in computer and information science up to this day brought us wirelessly connected small sized computers of which the most popular representative is the smartphone. It has become a ubiquitous and powerful device with far more processing power than even the rocket of the first successful moon mission had at its hand[1]. Circuit integration and microelectronic-mechanical systems (MEMS) led to sensor innovation and their implementation in the aforementioned mobile devices [Kar13]. They supported the user interaction (horizontal/vertical screen shift via tilt sensor, user-centered navigational map rotation, etc.) but as an additional result created new input options e.g. touch, speech, gesture. Those

[1] http://www.zmescience.com/research/technology/smartphone-power-compared-to-apollo-432/, last accessed03.01.2016

make the Human-Computer interaction more and more interactive even human-like leading to an interface which is more than ever natural. As shown in the past this attribute "natural" plays a big role in how fast HCI-systems are adopted by the user and hence determines its success and prevalence. With this in mind it is now the designers' intent to implement adaptive and intelligent natural user interfaces. The various sensors they employ for this provide for example visual and acoustic data about the nearby environment. Also motion and positional information can be captured. But these mostly unimodal sources are often relatively week in their content and usability. The compass for example can provide orientation. But in a standalone application it is not more than a nice gimmick. Adding additional information as GPS and map data enhances its usefulness, creating for example a user-orientation sensitive map navigation application [SMES14].
First research on the topic of advanced multimodal interfaces began to emerge as far back as 1980 [Bol80] but now with the already mentioned sensor integration in mobile devices this technology has a much greater purview and research on this topic is intensified (annual ACM meetings[2]: ICMI-MLMI, SIGDOC, IUI, SUI, UIST).
This research is motivated for several reasons. More complex information can be inferred combining the resources and forming multi-modal applications. Furthermore cross-modal integration leads to higher accuracy and error avoidance (e.g. accelerometer compensates drift of gyroscope)[3]. They provide alternative input channels (speech and/or gesture etc.) what can ease HCI for people which are impaired in one of the human interactive channels and most importantly people may process information faster and better when it is presented in multiple modalities [vWGP05] as also the human communication is inherently multimodal [Moe15].
Nevertheless the high information density has to be adequately conditioned before presented to the user and it has to be determined what is "natural"? For this reason researchers started to look at a system which is known to use multimodal information in the most natural way: the human brain. It is still a mystery how the human brain is able to process the high flow of information from the sensorial modalities of the body. But using neurophysiologic findings and concepts as a guideline to natural user interfaces may bring faster solutions reducing rejection probability and vice-versa will help to deepen the understanding of brain models in neuroscience.

According to this, in the following an introductory review is presented providing some key aspects for multimodal interfaces based on research in neuroscience. First a general model of the human multimodal integration is given and some essential brain mechanisms and restrictions are highlighted. Then the core design principles for a multimodal natural user interface in mobile devices are presented from the input over sensory data to high level information output. This document concludes stressing the main challenges and outlining the future of Natural User Interfaces.

[2] http://dl.acm.org/events.cfm, last accessed 03.01.2016

[3] http://www.invensense.com/products/motion-tracking/6-axis/mpu-6500/, last accessed 03.01.2016

Chapter 2

Cross-Modal Integration in Neuroscience

Brain recording technologies such as fMRI, EEG, MEG, single cell recordings and some others speeded up brain research significantly. Where previously only psychological methods could be used to investigate the human brain those technological tools enabled also neurophysiological examinations of active brain tissue. This led to the finding of functional zones as the visual, auditory, motor cortex etc. and first models of how the brain processes sensory input were created.
For the purpose of this paper research on general mechanisms and especially multimodal interactions in the brain was surveyed with two primary objectives. The first one is to infer a model for multimodal integration. Secondly attention was payed on the cognitive behavior of the brain in order to infer restrictions imposed by the mental capacity which should be taken into account for the design of a sophisticated user interface.

2.1 Brain Organization

Aside of the knowledge of brain areas it is a fact that information is transported via synapses which are interconnected with each other. Through this basic approach already the neurophysiological architecture of the brain indicates that information processing cannot be strictly unimodal. Similarly it is the length of those synaptic paths (can also be seen as an increased number of those interconnections) which elicits complex cognitive performance as an empiric measure for intelligence.
M.-M. Mesulman [Mes98] conducted an extensive survey on commonly accepted physiological principles that link sensation to behavioral outcome and presents us an exemplary model showing the interaction of visual and auditory neural processing (Fig.2.1).
Here the concentric rings represent various synaptic layers where the distance is given by the sequential response latency between specific functional zones or nodes symbolized by the small circles. On the same ring those areas are connected reciprocally.

The connections between the layers show the anatomical connections with monosynaptic distance from the primary sensory nodes up to trans-modal/shared zones. These pathways are bidirectional thus allowing feedback from higher nodes. While unimodal visual (green) and auditory (blue) pathways seem to be initially separate they both influence and are bound together by the hetero-modal/trans-modal areas. Equally lower nodes express low-level features such as color perception (V4) and movement perception (V5) but trans-modal nodes are critical for high-level information processing, e.g. from perception to recognition of faces or individuals. Those high-level nodes are not considered the origin of consciousness but gateways for the same (Cartesian Dualism). This is supported by experimental evidence. For example lesions of the connections to node f led to the prosopagnosia syndrome the inability to recognize faces. Mesulman states that this model leads to at least five networks which can be situated in the brain: Spatial-awareness; language; memory/emotion; face-object recognition; working memory. These networks share trans-modal nodes and therefore also the sensory modalities which are taste, smell, touch, audition, vision, proprioception, thermoception, nociception and equilibrioception.

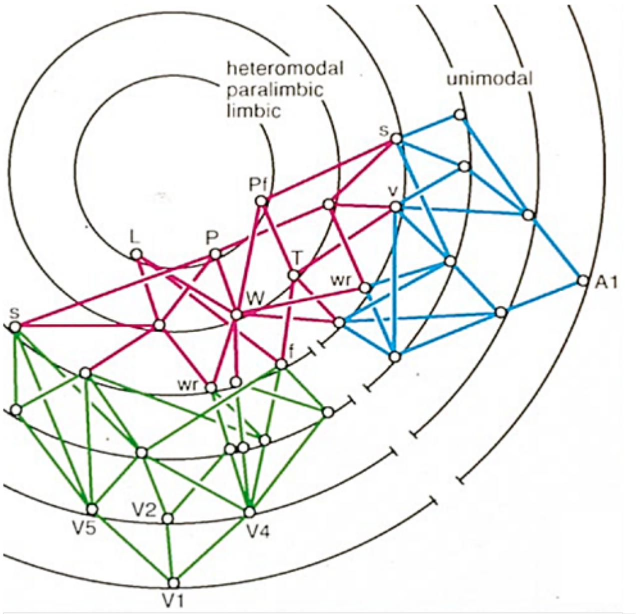

Figure 2.1: Physiological organization for sensorial integration according to Mesulman [Mes98]

2.2 Cross-modal Effects

Neurophysiological research showed trans-modal interaction exists. But the question whether the brain perceives information multimodal with a clear distinction between the senses or deals with content that is strongly dependent across the sensory modalities before perception happens cannot be answered by simple neuronal activity. There are several references in psychology indicating cross-modal influence:

Mc-Gurk effect: This perceptual phenomenon exemplifies that vision alters speech perception. The sound of /ba/ tends to be perceived as /da/ when it is coupled with a visual lip movement associated with /ga/. This happens also when the effect is known to the person under test beforehand. [McG76]

Ventriloquism: This effect shows how vision alters the spatial perception of sound. Presenting a test subject with a sound stimulus and simultaneously with a spatially disparate visual input the perceived location of the sound source shifts towards the location of the visual signal [HT66]. After training even without the visual stimulus the location of the sound remains shifted for a certain time in the tested environment. This is called the ventriloquist aftereffect and acts also as an example for neuronal plasticity.

Synesthesia: This is a mental condition in which a particular sensory event in one modality produces an additional sensory experience in a different modality. For example odor-taste synesthesia where odors e.g. vanilla are addressed to as sweet which is normally associated with taste stimulation [SPB99]. Another form is the perception of written words, numbers or letters as having colors [Cyt89]. The first example seems to apply to most of the population whereas most synesthesia effects are present in very few subjects.

Garner interference / Congruence effect: The failure to filter out irrelevant variation in an unattended perceptual dimension is known as the garner interference. Congruence effect is similar and refers to temporal performance variation when the stimulus has two or more perceptual normally polar attributes (high-low, bright-dark, etc.) that are congruent. These attributes can be from the same or different modalities. Garner interference generally decreases performance whereas congruence can have an enhancing effect. [Mar04]

Summing up the mentioned effects we can say cross-modal sensing exists and influences at least the spatial, temporal and perceptual performance of the brain. Nevertheless this cross-modal connection shows likewise advantages and disadvantages on the mental capability and when designing a system where cross-modality is implemented it has to be decided to what extent cross-modality should play a role and when.

2.3 Binding problem: early vs. late integration

This difficulty is commonly known as the *binding problem*. The dominant part of researchers advocates a post-perceptual or late integration. Among other reasons this is based on the cognitive development process. As high-level perception is not a prenatal property of the brain senses are said to be innately distinct modalities and are interrelated only through experience from interactions with the world. Some neurophysiological experiments incline also towards supporting the post-perceptual integration theory [Coe03].

This approach totally excludes the possibility of cross-modal influence before perception happens. However the effects discussed previously allow deducing early integration is indeed occurring. Nevertheless many implementations of multimodal interfaces still rely on late-integration with the simple reason of lower complexity. Training two channels separately takes less effort O(2N) than teaching the model two modes together O(N^2) [Tur13]. On the other hand early inducing of information across modalities could accelerate the outcome of the integration process and increase accuracy where unimodal processing alone wouldn't be successful.

2.4 Restraints in cognition

An important consequence of multimodal input and output in HCI- interaction is higher performance. But while we can increase artificially the output of computational systems the operation capacity of humans remains unchained.

Cognitive load: In the context of temporal critical tasks e.g. in cars the cognitive load plays a significant role. It has direct neural impact on the brain's working memory where new environmental and mental events are consciously processed [Mes98]. Investigating the source of cognitive load we differ two components: Intrinsic load which refers to the inherent complexity of the task itself and extraneous load which refers to the representational complexity of the task [SMP98]. Additionally the not task related load present in the mind adds to the mental burden. High cognitive load not only impairs the mental throughput but also learning and long memory storage is negatively affected. Moreover, once in the overload state, recovery time (Fig.2.1) is necessary despite the help of compensatory steps due to neuronal plasticity of the brain [CRC+11, Mes98].

But how is mental charge measurable in the first place? Applying cognitive load measurement for mobile interactive systems rules out an estimation method via self-report which lack real-time use. Physiological measures require mostly additional instruments (galvanic skin response). Chen et al. [CRC+11] promote measurements of performance or behavioral features. They carried out a representative study extracting fluctuations of speech and digital pen input.

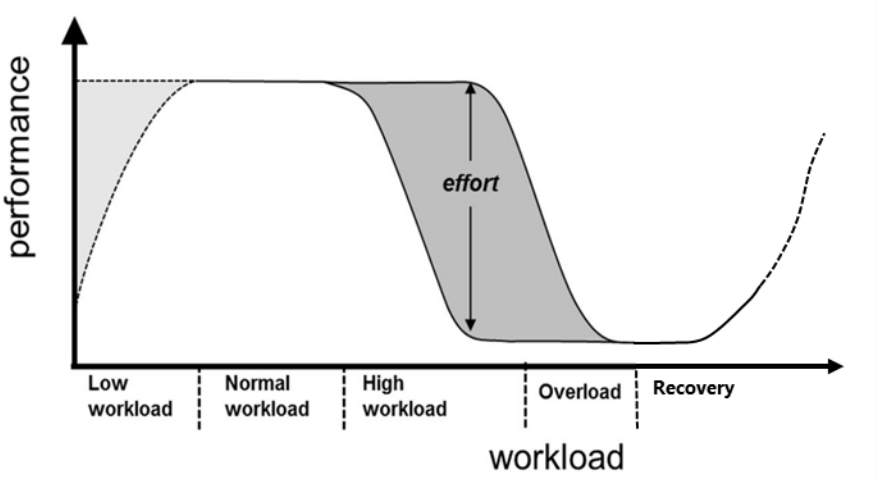

Figure 2.2: Relationship between performance and cognitive load, adapted from Chen [CRC+11]

Selective attention: Besides motivation and emotion, attention is substantial for modulation of sensory input. On the one hand it helps to filter relevant information by enhancing the selectivity, intensity and duration of a neuronal response [Mes98]. On the other hand it restricts the real-time processing capacity meaning that information gets lost as only a single channel of information can be attentionally processed at one time [Pas98]. The biological sensor architecture adapts to this. For example the most dominant sensor (80% of perceived information)[1] the eye uses the fovea to center focus on a restricted small region of the visual field of perception recognizing one object of attention. Research on visual selective attention showed performance increases when information is presented disparate in time (with at least 50 ms difference)[DD95]. Hence in interactive design the characteristic serially processing of attention has to be considered making sure the relevant information is conveyed properly.

Additionally there exists cross-modal dependency on non-attentional modalities. This can have an interfering effect as in the garner interference but with respect to the human evolution this clearly seems necessary. It would be a great safety issue if for example an auditory signal with high magnitude could not direct our attention forcefully. Nevertheless this mental condition could be exploited beneficially in interactive interface design directing the user's attention.

[1] http://simplybrainy.com/wp-content/uploads/2011/01/2008-Int-Vis-Other-Senses-All-Illustrations.pdf, last accessed 01.04.2016

Chapter 3

Multimodal Natural User Interface in mobile devices

Beyond the lessons in cross-modal design that can be derived from neuroscience the design of interfaces for mobile devices has to fit the technological framework given. For example the sensing modalities available are often quite different from the biological examples albeit the efforts to imitate or even surpass them. Appropriateness analysis is necessary to allocate their use for different tasks and their potential for multimodality. Looking at current technologies and applications some relevant implementation details and design principles stand out.

3.1 Common principles

A basic guideline for multimodal user interface design is defined by Reeves et al. [RLL$^+$04]. Summarizing their suggestions we can infer the keywords *Universality, Context awareness, Cognitive enhancement, Adaptability, Consistency, and Affordance*:

- **Universality:** Multimodal systems should be designed for the broadest range of users and contexts of use. Designers should support the best modality or combination of modalities anticipated in changing environments (for example office and driving a car).

- **Context awareness:** Designers should take care to address privacy and security issues in multimodal systems. For example, non-speech alternatives should be available in a public context to prevent others from overhearing provided information or conversations.

- **Cognitive Enhancement:** Maximize human cognitive and physical abilities, based on an understanding of users' human information processing abilities and limitations.

- **Adaptability:** Multimodal interfaces should adapt to the needs and abilities of different users, as well as different contexts of use. Individual differences (age, skill, sensory or motor impairment) can be captured in a user profile and used to determine interface settings. Furthermore modalities should be integrated in a manner compatible with user preferences and system functionality. For example, match the output to acceptable user input style.

- **Consistency:** Consistent sequences of actions should be required in similar situations and identical terminology should be used in system output, presentation and prompts, enabling shortcuts, state switching, etc.

- **Affordance:** Provide good error prevention and error handling e.g. reversal of actions; make functionality clear and easily discoverable.

The eight golden rules for user interface design by Shneiderman [SP10] mostly confirm this guideline but some principles are emphasized and other concepts are added: Another central idea is supporting the *"internal locus of control"* meaning the user has a feeling of power or control over the system. Providing shortcuts or alternative modality use for (experienced) users is one way to do so.
Unfortunately an increase in alternative control mechanisms would require additional learning. In order to reduce the effort for the user the system should be communicative and provide informative feedback. This could simply mean an answer-back signal which confirms an active recording of input (e.g. keyboard-click) or even an affordance language which helps to explore the possible ways of input via evocative feedback at least for unexperienced users. In multi-touch control for example a visual response to an initial input (one finger, multiple fingers or palm) could call a shadow guide indicating possible follow-up motions in order to perform different kinds of commands for selection operation or object manipulation [FBRMW09].
Some more aspects should be kept in mind as well. Based on different criteria the relevant input or output modalities for a certain use case or situation have to be picked. Also the *expressiveness* saying how well a modality transfers the information for a specific task should be regarded [Rat08]. Furthermore the *social acceptability* influences this choice. In a conference the output or input of a mobile phone being auditory and not visual is not acceptable. When it comes to gesture input a beforehand seldom integrated factor might be c*ultural differences.* Hand gestures could convey different information and might even be offensive depending on the cultural background.
Therefore a thorough task analysis before implementation of different modalities is recommended considering the mentioned principles. But also the system itself should automatically adapt to different situations depending on the context using so called tasking applications [Moe15] or inherently built-in software in real-time. The task analysis in the form of evaluation is discussed in chapter3.4.1. A short discussion of the term context follows in the next part.

3.2 Role of context

A prominent and reverberating term in intelligent system design is context. Context information is the necessary outcome of the multimodal sensor integration in interfaces in order to act according to the mentioned principles. In neuroscience context represents high-level perceptual information steering the conscious behavioral output. The awareness of context is considered a key enabler for next generation information services [Sch05]. Context is not defined strictly but it is commonly referred to as information that can be used to characterize the situation of an entity [Dey01]. The philosopher John Dewey discerns two categories interpreted as "contemporary"/"spatial"/"extrinsic" and "background"/ "temporal"/ "intrinsic" context, where contemporary parameters are measurable and the intrinsic part is abstract and relates to tradition/culture, mental habits, experience, etc., as in figure3.1[EM01]. In computer science a situation is often categorized by location, time, identity, and activity answering the questions where, when, who and what [Kar13, SMES14]. It's important to note that these categories of information should not be quantitative but qualitative similar to human awareness of context. If given a context description task e.g. whether we go to bed is not determined by the time of the day but rather by the daytime i.e. morning or evening. Moreover this example shows that time is not enough to contextualize a situation. The information about location or habit could alter the context significantly as it is unlikely that the user is sleeping while not being indoors. Nevertheless this kind of context awareness for natural interfaces is not yet totally complete as it lacks the subjective valuation and weighting existing in human-human interaction due to empathy, mood, attention and habits. But as this most often contradicts a purely logic methodology of computer design, fully including intrinsic context would increase a systems complexity significantly. Therefore the proposed four attribute paradigm serves the purpose for a first context implementation in a user interface.

With this also an enhanced implicit or passive interaction is possible without deliberate user influence. This goes beyond traditional context awareness referring to an active user interaction tracking and relates more to a context-driven or sensor-driven interaction [Moe15, LO12].Including context is actively addressed in research for more sophisticated and intelligent interface designs for context-aware or context-driven systems with augmented cognition.

3.3 Multimodal fusion

Originally input in interactive systems was solely unimodal (GUI) whereas output was already multimodal (graphics, sound, etc.). Nowadays in devices like smartphones and tablets various input modalities are added (speech, multi-touch, gestures, motion, etc.). Those are "natural (but ambiguous) inputs, such as speech, and less natural (but unambiguous) inputs by way of direct manipulation" [Rou10]. Their integration or fusion will allow interpretation for the appropriate task in a certain

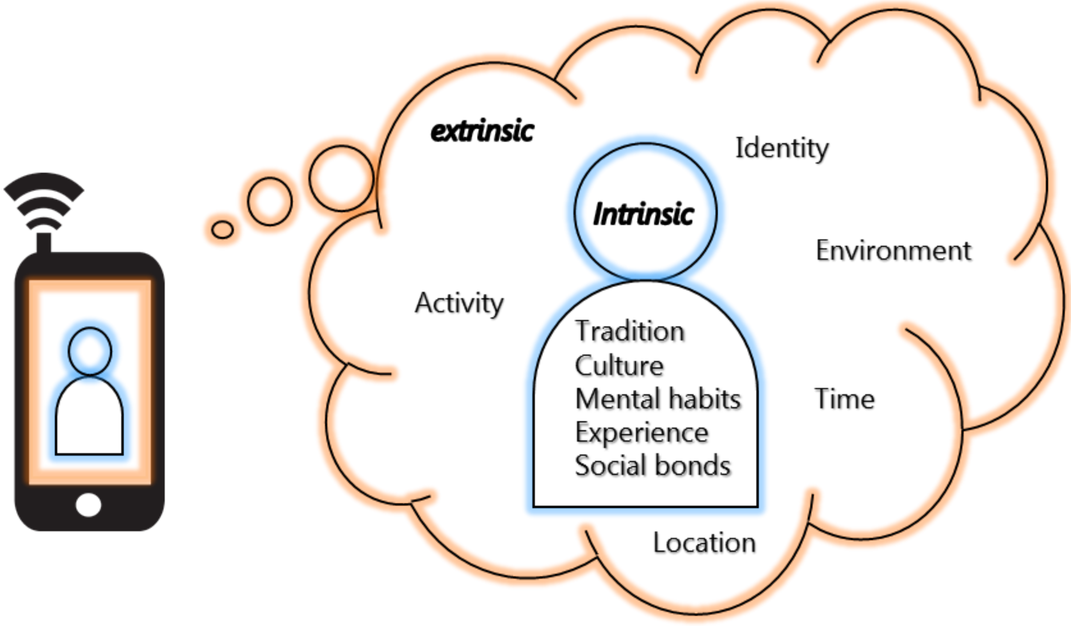

Figure 3.1: Examples of extrinsic and intrinsic context information for mobile context-aware systems

context. According to Nigay and Coutaz [NC09] multimodal input combination can be categorized in 4 ways: **alternative, exclusive, concurrent and synergistic** (Tab.3.1). Additionally to the type of fusion (late/independent or early/combined) this classification scheme considers the manner of input by the user (sequential or parallel). For example when different modalities are used at the same time and the multimodal sensory information is treated as a whole the system is synergistic.

Table 3.1: Design space for multimodal systems adopted by Nigay and Coutaz[NC09]

		Use of modalities	
		Sequential	Parallel
Fusion	Combined	**Alternative**	**Synergistic**
	Independent	**Exclusive**	**Concurrent**

Similar to the question of early or late integration in neuroscience fusion has to be performed at some point in the integration process. This can happen early at the **feature level** (multimodal information is present in form of patterns or characteristics) or late at the **decision level** (rules and constraints). The sensor level acts as third existing layer for fusion but the raw data information is seldom suitable because this step is depending mostly on the used sensor technology. Generally various fusion methods of table 3.1 can be used at different levels and on various features within one sensor modality or across various modalities.

Speech for example can be expressed in many ways. Even in one language the

various pronunciations, dialects, background noise and individual characteristics of the vocal cord make uniformity in voice commands impossible. The same applies to other interaction modalities like gesture or facial expressions. The input therefore is non-deterministic. Probabilistic classifiers have to extract the important features from the raw data (Hidden Markov Model HMM, Gaussian Macro Model GMM, Self-Organizing Map SOM, etc.). In computer vision characteristic properties of images are extracted with SIFT, MSER, SURF, FAST, BRIEF, or ORB algorithms [Moe15]. After this step the results are only valid up to a certain degree. Aside of the non-deterministic input the unpredictable dynamics of modality combinations with varying temporal distance complicate the right interpretation. But on the other hand the existing redundancy in some multimodal inputs bolsters specific decisional output. Additionally considering the task, context and user preference detection could narrow down the meaning significantly. However this type of information as well is mostly non-deterministic and susceptible to erroneous interpretation. In human-human interaction this is resolved by feedback in both directions in the form of active conversation. In consequence an interactive system should provide similar mechanisms [LNP+09].
Manifold toolkits and frameworks for context aware applications, programming multimodal behavior or modeling multimodal interactions exist. Those facilitate the application development process and are important for a fast time to market and new innovations. Presenting them in detail would exceed this works scope. A comparative and comprehensive review is given by Dumas et al.[DLO09] and Möller [Moe15] respectively.

3.4 Mobile device principles

In this work we want to focus on natural interfaces on mobile devices which are a central part for the research on ubiquitous computing. We carry them around all day and use them for all sorts of things: entertainment, communication, information procurement. The progressing sensor integration and their installation in mobile devices make them able to measure our environment in many ways. Beyond their traditional utilization as e.g. for device control, data capture or gaming applications their context detection potential is great:

- For *cognitive load* auditory information from the microphone combined with pen input or pressure sensor data form the touch sensors could be used [CRC+11].
- User *identity* could be detected using the front RGB camera or special infrared cameras.
- *User or device activity* is a possible result from integration of motion sensor data (Accelerometer, Pedometer, etc.), camera information or even physiological sensors (heart rate, transpiration).

- *Location* (indoor, outdoor, position) could be determined using communication sensors as WIFI or GPS and the camera data for example.

Here we can generally differ between physical context sensors - measuring gravity, temperature etc. - and virtual context sensors. The latter refer to the information we can infer from interpersonal communication as relationship status e.g. if someone is a friend or colleague. This encloses the field of social context.
Hence there is a vast amount of sensors available and this makes mobile devices first choice for an implementation of multimodal context-aware interfaces although hitherto there is only a sparse exploitation of multiple sensors [Moe15]. But their potential for new application areas is great. The main profiting areas involving natural user interfaces and multimodal integration in mobile devices at the time are at first activity recognition mostly for health tracking or for frailty warning systems which call an ambulance if an elderly person is falling severely [GvdVN14]. Some effort is also put into systems supporting multimodal learning as knowledge acquisition works best via several stimulated channels. Furthermore there is a clear usability for Pedestrian Navigation Systems (PNS) especially for indoor navigation where higher localization accuracy is needed. There are manifold realizations e.g. map-based interfaces or camera views with superimposed virtual elements (augmented reality). But the person who decides at the end which implementation is the best is the user.

3.4.1 User centered design and evaluation

Therefore it is crucial that the designer involves the end user very early in the design process and development. For this purpose many evaluation methods exist from surveys over evaluation in the laboratory to data collection in the field. Mostly it depends on the implementation progress which category of methods is applicable and their outcome differs greatly with different levels of user experience. Evaluation is an archetypal part of a development process and with focus on mobile multimodal interface design Andreas Möller [Moe15] proposes an iterative -case study probed-evaluation process addressing all of the main categories. Initially questionnaires help to fathom the needs of users and give first feedback. An interactive discussion with a focus group might elicit some sticking points where special care should be put in. For more quantitative data, laboratory and real-world evaluation is necessary. Is the implementation of the multimodal interaction not mature enough the Wizard-of-Oz approach (WOz) is often used, where the researcher is responsible for the output of the system or, in other words, can help out where the system would yet fail [Moe15]. Finally first prototypes should be tested in the field in order to test the direct impact of the design and to proof the concept (Tab.3.2).

Table 3.2: Iterative design model of a multimodal system for mobile devices adapted from Möller[Moe15]

Method	***Focus of Interest***	***Why Important?***
Focus Group	Initial insights on user needs	• Fast • Inspiring • Directions for research questions
Large-Scale Questionnaire	Broad feedback on concept or mockup	• Early feedback prior to implementation • Heterogeneous participants
Laboratory Evaluation: Wizard of Oz	Quantitative and qualitative handson feedback	• User experience with prototype • Controlled conditions for measurements • Face-to-face interaction with participants
Real-World Evaluation	Usage and adoption in the field	• Relation to context and environment • Degree of novelty • Long-term usage

3.4.2 Battery life

In mobile device design one has to consider a restricted amount of resources [LO12]. On the technical side the size is a factor but the biggest hardware constraint also known to the user is the battery life. Multimodality increases the amount of active current sinks as a varying number of multiple sensors are used. Additionally for high-level interaction with context-aware systems a constant tracking of information has to be performed where the sensors remain in an always-on state [Kev15]. The main power draining hardware is the display, radios and the application processor. For this reason technology providers nowadays are directly implementing sensor hubs where the sensors have a specifically built low-power processor nearby which is specialized and optimized for data acquisition or integration. With this co-processor approach, parallel processing and direct hardware acceleration is supported speeding up the device, and low power listening modes can be implemented e.g. Audience's VoiceQ technology in the N100 Natural User Experience Multisensory processor[1]. This single chip provides not only automated hot word detection for device interaction but incorporates also motion sensor fusion of an inert motion unit (IMU) and pressure data presenting the application interface designer with qualitative context information of user's posture, location, activity and mode of transport.

[1] http://audience.com/nue-n100-video, last accessed 04.01.2016

3.4.3 Data management

With the increasing amount of data another restriction becomes clear. Data storage capacity is a matter of size and costs. Therefore especially mobile devices need a good data management for tracking all the sensor data. The next question is at which level of data fusion the storage should be performed. High level data might be very task specific but might be applicable for different applications whereas low level data is raw and big while being redundant and noisy. The task of the device goes beyond data management and knowledge management would be a more appropriate term. Radio provides the devices with a bypass to this problem.

It connects several devices making each individual device more performant if e.g. data can be directly accessed and has not to be measured first. Researchers actively address the implementation of multi-device interfaces connected via multiple wireless channels as near field radio (RFID, Bluetooth, NFC, etc.) or far distance radio (WIFI, GSM, UMTS, LTE, etc.) [Rou10].

Furthermore it connects the device to the Internet and its almost unlimited resources. But although Server storage is not limited in capacity, connectivity is still frequently interrupted with mobile devices. A smart trade-off between offline data and online data is necessary e.g. when detecting a tunnel on the route the device will temporarily download the map data for offline availability(Google Maps[2]).

Hence server storage is a powerful tool and should be considered for a supporting and seamless interface interaction and seems only natural taking into account the advent of the "Internet of Things"in our current world of "Big Data" [Kar13].

3.4.4 Multimodal interaction

A natural user interface for ubiquitous devices expresses itself in a continuous adaption of human-computer interaction at run-time according to user, device, and environmental context supporting multiple interaction modalities, channels, and devices. This should be done by the system choosing an active interaction model out of a set of various interaction models. This constitutes a reasonable constraint as humans do similar in conversations. They are limiting the area of discourse for less mental demand while still being able to dynamically change the interaction mode [Blu10].

We comprehensively discussed the distinction between multimodality and cross-modality in neuroscience. Talking about the implementation of a natural user interface we only used the term multimodal. On a system level cross-modality is only present in a sense where sensors are inherently combined for higher accuracy and robustness as in an inert motion unit for example. But the output of an HCI-system has to take into account the cross-modality of the brain and other mental constraints in order to make the interaction pleasant and feeling natural.

[2] http://t3n.de/news/google-maps-offline-navigation-655065/, last accessed 04.01.2016

Chapter 4

Outlook

Mobile devices already form a crucial part in the new world of ubiquitous computing and will be key-nodes for controlling and communication in the upcoming "Internet of Things". Next steps will be the interface migration or distribution over several devices [Blu10], exploring a wider range of modality combinations using more than two modalities, and the investigation of new modalities for example weight shifting, shape changing or ambient life-like actuation as breathing [KHR10]. The longer vision are "butler-like" interfaces which are geared to a more human form of communication not only acting according to context but anticipating it and understanding the user's feelings and idiosyncrasies [Tur13]. But the road heretofore will be rocky and full of challenges.
If it comes to the direct implementation there is no established standard guideline used by the majority of the designers. There is also a lack of standardization across various operating systems[Kar13]. Furthermore each technology is an active research area in itself be it speech/vision-based recognition, integration, or machine learning techniques and user related problems and constraints as the cognitive load have to be understood before overshooting the capabilities of the human brain. Similarly context information has no generic computational representation and appropriate user intent assignment and context extraction remains the most complex part of the design process. With the increased connectivity of multiple devices and the server-client based data storage becoming more and more popular those systems also are very affine to security breaches[Tur13].Finally it will always be difficult for the user and the designer as well to move beyond known principles of HCI in order to create a Natural User Interface. We may be in a post-WIMP world, but still prefer the traditional interaction when it comes to complex and demanding tasks. The challenge is to create interfaces where interaction is feeling natural which adapt themselves smoothly and smartly, which are not relying on a user's intuition but are anticipatory and evocative. The interaction should be seamless, synchronized, and extensive in its possibilities leading to new improved forms of NUIs as for example organic user interfaces (OUI) (Wixon 2008)[1]

[1] UX Week 2008,Dennis Wixon https://vimeo.com/2893051, last accessed 04.01.2016

List of Figures

Bibliography

[Blu10] Marco Blumendorf. Ubiquitous user interfaces: Multimodal adaptive interaction for smart environments. In Stan Kurkovsky, editor, *Multimodality in Mobile Computing and Mobile Devices: Methods for Adaptable Usability*, chapter 2, pages 24–52. IGI Global, 2010.

[Bol80] R.A. Bolt. Put-that-there: voice and gesture at the graphics interface. In *ACM Comput. Graphic.*, volume 14, pages 262–270, 1980.

[Coe03] M.H. Coen. Multimodal integration: a biological view. In *. Int. Joint Conf. Artif. Intell.*, volume 17, pages 1417–1424, 2003.

[CRC$^+$11] F. Chen, N. Ruiz, E. Choi, J. Epps, A. Khawaja, R. Taib, B. Yin, and Y. Wang. Multimodal behavior and interaction as indicators of cognitive load. In *ACM Trans. Interact. Intell. Syst.*, volume 2, pages 39–72, 2011.

[Cyt89] R. E. Cytowic. *Synaesthesia: A union of the senses*. MIT Press, 1989.

[DD95] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

[Dey01] A.K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5:4–7, 2001.

[DLO09] B. Dumas, D. Lalanne, and S. Oviatt. Multimodal interfaces: a survey of principles, models and frameworks. In *Human Machine Interaction. Lecture Notes in Computer Science*, volume 5440, pages 3–26, 2009.

[EM01] H.R. Ekbia and A.G. Maguitman. Context and relevance: A pragmatic approach. In *Lecture Notes in Computer Science*, volume 2116, pages 156–169, 2001.

[FBRMW09] D. Freeman, H. Benko, M. Ringel Morris, and D. Wigdor. Shadowguides: Visualizations for in-situ learning of multi-touch and whole-hand gestures. In *Proceedings of Tabletop*, pages 183–190, 2009.

[GvdVN14] J.J. Guiry, P. van de Ven, and J. Nelson. Multi-sensor fusion for enhanced contextual awareness of everyday activities with ubiquitous devices. *Sensors*, 14:5687–5701, 2014.

[HT66] I. P. Howard and W. B. Templeton. *Human spatial orientation*. Wiley, 1966.

[Kar13] Kaivan Karimi. The Role of Sensor Fusion and Remote Emotive Computing (REC) in the Internet of Things, 2013.

[Kev15] D.T. Kevan. Sensor Fusion Growth Creates New Challenges, 2015.

[KHR10] S. Kratz, F. Hemmert, and M. Rohs. Natural user interfaces in mobile phone interaction. In *ACM CHI*, 2010.

[LNP$^+$09] D. Lalanne, L. Nigay, P. Palanque, P. Robinson, J. Vancerdonckt, and J-F. Ladry. Fusion engines for multimodal input: A survey. In *ACM ICMI-MLMI*, pages 153–160, 2009.

[LO12] Tom Lovett and Eamonn O'Neill. *Human spatial orientation*. Springer London, 2012.

[Mar04] L.E. Marks. Cross-modal interactions in speeded classification. In G.A. Calvert, Ch. Spence, and B.E. Stein, editors, *The Handbook of Multisensory Processes*, chapter 6, pages 85–105. MIT Press, 2004.

[McG76] J. McGurk, H.and MacDonald. Hearing lips and seeing voices. *Nature*, 1:746–748, 1976.

[Mes98] M.-Marsel Mesulman. From sensation to cognition. *Brain*, 121:1013–1052, 1998.

[Moe15] Andreas Moeller. *Leveraging Mobile Interaction with Multimodal and Sensor-Driven User Interfaces*. PhD thesis, Technical University of Munich- Department of Electrical Engineering and Information Technology, 2015.

[NC09] L. Nigay and J. Coutaz. Fusion engines for multimodal input: A survey. In *ACM ICMI-MLMI*, pages 153–160, 2009.

[Pas98] H. E. Pashler. *The psychology of attention*. MIT Press, 1998.

[Rat08] A. Ratzka. Context and relevance: A pragmatic approach. In *Engineering Interactive Systems Lecture Notes in Computer Science*, volume 5247, pages 58–71. Springer Berlin Heidelberg, 2008.

[RLL+04] L.M. Reeves, J. Lai, J.A. Larson, S. Oviatt, T.S. Balaji, S. Buisine, P. Collings, P. Cohen, B. Kraal, J.-C. Martin, M. McTear, T.V. Raman, K.M. Stanney, H. Su, and Q.Y. Wang. Guidelines for multimodal user interface design. In *ACM*, volume 47, pages 57–59, 2004.

[Rou10] Jose Rouillard. Multimodal and multichannel issues in pervasive and ubiquitous computing. In Stan Kurkovsky, editor, *Multimodality in Mobile Computing and Mobile Devices: Methods for Adaptable Usability*, chapter 1, pages 1–23. IGI Global, 2010.

[Sch05] A. Schmidt. A layered model for user context management with controlled aging and imperfection handling. In *Modelling and Retrieval of context Lecture Notes in Artificial Intelligence*, volume 3946, pages 86–100. Springer Berlin Heidelberg, 2005.

[SMES14] S. Saeedi, A. Moussa, and N. El-Sheimy. Context-aware personal navigation using embedded sensor fusion in smartphones. *Sensors*, 14(4):5742–5767, 2014.

[SMP98] J. Sweller, J. Merrienboer, and F. Paas. Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3):251–296, 1998.

[SP10] B. Shneiderman and C. Plaisant. *Designing the User Interface: Strategies for Effective.* Addison-Wesley Publ. Co., 5. edition, 2010.

[SPB99] R. J. Stevenson, J. Prescott, and R. A. Boakes. Confusing tastes and smells: How odors can influence the perception of sweet and sour tastes. *Chemical Senses*, 24:627–635, 1999.

[Tur13] Matthew Turk. Multimodal interaction. A review. *Pattern Recognition Letters*, 36:189–195, 2013.

[vWGP05] V. van Wassenhove, K.W. Grant, and D. Poeppel. Visual speech speeds up the neural processing of auditory speech. In *Proceedings of the Academy of Natural Science*, volume 2005, pages 1181–1186, 2005.

License