Proceedings of the 1st Joint Workshop on Smart Connected and Wearable Things 2016

Co-located with 21th ACM Conference on Intelligent User Interfaces (ACM IUI 2016)

Dirk Schnelle-Walka, Lior Limonad, Florian Müller, Joel Lanir, Tobias Grosse-Puppendahl;
Massimo Mecella, Kris Luyten, Tsvi Kuflik, Oliver Brdiczka, Max Mühlhäuser

TECHNISCHE UNIVERSITÄT DARMSTADT

Contents

1 Preface

There is an undeniable ongoing trend to put computing capabilities into everyday objects and onto body-worn devices. Examples range from smart kitchen appliances (smart coffee machines, smart knifes and cutting boards), fitness trackers, smart glasses up to smart meeting rooms and even urban infrastructures.

The SCWT workshop integrated the former workshops on Interacting with Smart Objects (SmartObjects[1]) and the Internet of Wearable Things (IoWT[2]). Therefore, the workshop topics focused on advanced interactions with smart objects in the context of the Internet-of-Things (IoT), and on the increasing popularity of wearables as advanced means to facilitate such interactions. The combined focal point of the workshop was a good fit with the IUI conferences overall aim to blend between HCI and the AI communities, and also between academia and industry. Consequently, the workshop brought together researchers and practitioners interested in IoT and wearable-computing based applications, including also the interesting interplay between this emerging field and the more mature areas of mobile and ubiquitous computing.

1.1 Introduction

From the perspective of the smart objects, many of the objects are functional on their own, but added value is obtained through communication and distributed reasoning. While other venues have focused on the many technical challenges of implementing smart objects, far less research has been done on the topic of how the intelligence situated in these smart objects can be applied to improve their interaction with the users. This field of study poses unique challenges and opportunities for designing smart interaction.

From the perspective of wearables, the enablement of new wearable-based solutions requires synergy among multiple disciplines, such as Machine Learning, Signal Processing, Human-Computer Interaction, Data and Knowledge Representation, Information Visualization, Computational Neurosciences, and even Education. For example, there are many implications to mental and emotional signals that can be revealed via the use of wearables (e.g., EEG, heart-rate variability, GSR) as a means to determine whether the user is capable of performing her tasks, let it be a field worker, a combat pilot, a stock broker, or an elder user living alone. The blending of smart objects and wearables enables the creation of new solutions that combine data of multiple sensors and data stores as the basis for a more intimate and timely interaction with things and users, and high level of situational awareness. Special attention may be given to the creation of innovative solutions, not only applicable to consumers, but also to the enterprise. This can possibly be the basis for newly created business services.

1.2 Workshop Format

The SCWT'2016 workshop was held co-located with the 21st ACM IUI 2016 conference (IUI 2016), March 7-10 2016, Sonoma, California. The workshop combined the two previously isolated workshops SmartObjects and IoWT. Our full-day workshop accepted submissions in the following three categories:

- position papers and posters (2 pages) focusing on novel concepts or works in progress,
- demo submissions (2 pages) and
- full papers (4-6 pages) covering a finished piece of novel research.

Our goal was to attract high-quality submissions from several research disciplines to encourage and shape the discussion, thus, advancing the research of interacting with smart objects and wearables. This strategy lead to a lively and productive discussion during the workshop. We also summarized the outcome and published it on the workshops website in addition to the joint TU prints proceedings. This publication strategy attracted higher quality submissions, and increased the exposure of the workshop before and after the event.

The combination of the SmartObjects and the IoWT workshops lead to interesting multi-disciplinary discussions and proved to be a valuable exchange of ideas.

[1] http://www.smart-objects.org/
[2] https://sites.google.com/site/iowt2016/

1.3 Workshop Committee

Program Co-Chairs (Organizers)

Interacting with Smart Objects:

- Dirk Schnelle-Walka, Harman International, Germany
- Florian Müller, TU Darmstadt, Germany
- Tobias Grosse-Puppendahl, Microsoft Research, United Kingdom
- Benedikt Schmidt, TU Darmstadt, Germany
- Kris Luyten, iMinds - Expertise Centre for Digital Media - Hasselt University, Belgium
- Oliver Brdiczka, Vectra Networks, USA
- Max Mühlhäuser, TU Darmstadt, Germany

Internet of Wearable Things:

- Lior Limonad (liorli@il.ibm.com), IBM Research
- Joel Lanir (ylanir@is.haifa.ac.il), University of Haifa
- Massimo Mecella (mecella@dis.uniroma1.it), Sapienza University of Rome
- Tsvi Kuflik (tsvikak@is.haifa.ac.il), University of Haifa

Program Committee

Interacting with Smart Objects:

- Bo Begole, Samsung, USA
- Marco Blumendorf, smartB, Germany
- Aba-Sah Dadzie, Open University, UK
- Fahim Kawsar, Bell Labs, Belgium
- Alexander Kröner, Technische Hochschule Nürnberg, Germany
- Germán Montoro, UAM Madrid, Spain
- Patrick Reignier, Inria, France
- Geert Vanderhulst, Bell Labs, Belgium
- Raphael Wimmer, Universität Regensburg
- Massimo Zancanaro, FBK, Italy
- Le T. Nguyen, Carnegie Mellon University, Mountain View, CA, US

Internet of Wearable Things:

- Sourav Bhattacharya, Bell labs
- George DeCandio, IBM Research
- Marlon Dumas, University of Tartu

- Dan Feldman, University of Haifa
- Antonio Kruger, Saarland Univesrity and DFKI
- Mudhakar Srivatsa, IBM Research
- Anthony Tang, University of Calgary

Exploring the potential contribution of mobile eye-tracking technology in enhancing the museum visit experience

Moayad Mokatren
The University of Haifa
Mount Carmel
moayad.mokatren@gmail.com

Tsvi Kuflik
The University of Haifa
Mount Carmel
tsvikak@is.haifa.ac.il

ABSTRACT
In many cases, visitors enter the museums with guide book or with a device that supports delivering content about the exhibits. Studies have shown that such content delivery has the potential of enhancing the museum visit experience. Earlier studies relied on using sensors for location-awareness and interest detection. This work-in-progress explores the potential of mobile eye-tracking technology in enhancing the museum visit experience by building and developing a system that runs on handheld device and uses Pupil-Dev mobile eye tracker. Our hypothesis is that the use of the eye tracking technology in museums' mobile guides can enhance the visit experience by enabling more intuitive interaction. We got satisfactory preliminary results when we examined the performance of a mobile eye tracker in a realistic setting. Future work will focus on building and experimenting a complete system in a realistic setting.

Author Keywords
Mobile guide; Mobile eye tracking; Personalized information; Smart environment; Context aware service.

ACM Classification Keywords
H.5.2. Input devices and strategies (e.g., mouse, touchscreen)

1. INTRODUCTION

The museum visit experience has been changed over the last two decades. With the progress of technology and the spread of handheld devices, many systems were developed to support the museum visitor and enhance the museum visit experience. The purpose of such systems was to encourage the visitors to use devices that provide multimedia content rather than use guide books, and as a consequence focus of the exhibits instead of flipping through pages in the guide book [Ardissono et al. 2012; Stephens 2010; Billings 2009; Cheverst et al. 2000].

Understanding the museum visitors' motivations plays a crucial role in the development and designing of systems that support their needs and could enhance their visit experience. Falk and Dierking [2000] and Falk [2009] tried to answer the question of what do visitors remember from their visit and what factors seemed to most contribute to visitors' forming of long-terms memories: "when people are asked to recall their museum experiences, whether a day or two later or after twenty or thirty years, the most frequently recalled and persistent aspects relate to the physical context-memories of what they saw, what they did, and how they felt about these experiences.". Zancanaro et al. [2009] that followed Veron and levassier [1983], and Dim and Kuflik [2014] explored the potential of novel, mobile technology in identifying visitors behavior types in order to consider what/how/when to provide them with relevant services. A key challenge in using mobile technology for supporting museum visitors' is figuring out what they are interested in. One key aspect is tracking where the visitors are and the time they spend there [Yalowitz and Bronnenkant, 2009]. A more challenging aspect is finding out what exactly they are looking at [Falk and Dierking, 2000]. Lanir et al. [2013] tried to examine the influence of a location-aware mobile guide on museums visitors' behavior, by comparing behavior of visitors who used a mobile multimedia location-aware guide during their visit and that of visitors who did not use any electronic aid. Their results indicate that visitors' behavior was altered considerably when using a mobile guide. Visitors using a mobile guide stayed at the museum longer and were attracted to and spent more time at exhibits where they could get information about.

What is usually common to most context aware services nowadays is that they make use of the communication and computational power (and sensors) of the users' mobile devices (e.g. mostly smartphones). In addition, they interact with their users mainly by their mobile device's touch screens, which has one major limitation: the users have to look at them during the interaction (even though, voice commands can be performed with applications like SIRI, this option is still very limited). Moreover, a major challenge in this aspect is to know exactly what the user is interested in. "A central problem in location-aware computing is the determination of physical location. Researchers in academia and industry have created numerous location-sensing systems that differ with respect to accuracy, coverage, frequency of location updates, and cost of installation and maintenance." [Hazas et al, 2004].

Given todays' performance of our mobile devices, we should be able to gain access seamlessly to information of interest, without the need to take pictures or submit queries and look for results, which are the prevailing interaction methods with our mobile devices. As we move towards "Cognition-aware computing" [Bulling and Zander 2014], it becomes clearer that eye-gaze based interaction should and will play a major role in human-computer interaction before/until brain computer interaction methods will become a reality [Bulling et al. 2012].

The study of eye movements started long time ago (almost 100 years ago), Jacob and Karn [2003] presented a brief history of techniques that were used to detect eye movements, the major works dealt with usability researches, one of the important

works started in 1947 by Fitts and his colleagues [Fitts, Jones & Milton, 1950] when they began using motion picture cameras to study the movements of pilots' eyes as they used cockpit control and instruments to land an airplane. "It is clear that the concept of using eye tracking to shed light on usability issues has been around since before computer interfaces, as we know them" [Jacob and Karn 2003]. Certain mobile eye tracking devices that enables to detect what someone is looking at and stores the data for later use and analysis, have been developed and could be found in the market nowadays [Hendrickson et al. 2014].

In recent years, eye tracking and image based object recognition technology have reached a reliable degree of maturity that can be used for developing a system based on it, precisely identifying what the user is looking at [Kassner et al. 2014]. We shall refer to this field by reviewing techniques for image matching and extend them for location-awareness use, we will follow the approach of "What you look at is what you get" [Jacob 1991].

With the advent of mobile and ubiquitous computing, it is time to explore the potential of this technology for natural, intelligent interaction of users with their smart environment, not only in specific tasks and uses, but for a more ambitious goal of integrating eye tracking into the process of inferring mobile users' interests and preferences for providing them with relevant services and enhancing their user models, an area that received little attention so far. This work aims at exploring the potential of mobile eye tracking technology in enhancing the museum visit experience by integrating and extending these technologies into a mobile museum visitors' guide system, so to enable using machine vision for identifying visitors' position and identifying the object of interest in this place, as a trigger for personalized information delivery.

2. BACKGROUND

2.1 Museum visitors and their visit experience

Understanding who visits the museum, their behaviors and the goal of the visit can play an important role in the design of museums' mobile guide that enhances the visit experience, "the visitors' social context has an impact on their museum visit experience. Knowing the social context may allow a system to provide socially aware services to the visitors." [Bitgood 2002; Falk 2009; Falk and Dierking 2000; Leinhardt and Knutson 2004; McManus 1991; Packer and Ballantyne 2005].

Falk [2009] argued that many researches have been done on who visits museums, what visitors do in the museum and what visitors learn from the museum, and tried to understand the whole visitor and the whole visit experience as well as after the visit. Furthermore, he proposed the idea of visitors "identity" and identified five, distinct, identity-related categories:

- *Explorers: Visitors who are curiosity-driven with a generic interest in the content of the museum. They expect to find something that will grab their attention and fuel their learning.*
- *Facilitators: Visitors who are socially motivated. Their visit is focused on primarily enabling the experience and learning of others in their accompanying social group.*
- *Professional/Hobbyists: Visitors who feel a close tie between the museum content and their professional or hobbyist passions. Their visits are typically motivated by a desire to satisfy a specific content-related objective.*
- *Experience Seekers: Visitors who are motivated to visit because they perceive the museum as an important destination. Their satisfaction primarily derives from the mere fact of having 'been there and done that'.*
- *Rechargers: Visitors who are primarily seeking to have a contemplative, spiritual and/or restorative experience. They see the museum as a refuge from the work-a-day world or as a confirmation of their religious beliefs.*

In addition, he argued that the actual museum visit experience is strongly shaped by the needs of the visitor's identity-related visit motivations, and the individual's entering motivations creates a basic trajectory for the visit, though the specifics if what the visitor actually sees and does are strongly influenced by the factors described by the Contextual Model of Learning:

- Personal Context: The visitor's prior knowledge, experience, and interest.
- Physical Context: The specifics of the exhibitions, programs, objects, and labels they encounter.
- Socio-cultural Context: The within-and between-group interactions that occur while in the museum and the visitor's cultural experiences and values.

Nevertheless the visitor perceives his or her museum experience to be satisfying if this marriage of perceived identity-related needs and museum affordance proves to be well-matched. Hence, considering the use of technology for supporting visitors and enhancing the museum visit experience, it seems that these aspects need to be addressed by identifying visitors' identity and providing them relevant support.

2.2 Object recognition and image matching

Modern eye trackers usually record video by a front camera of the scenes for further analysis [Kassner et al. 2014]. Object recognition is a task within computer vision of finding and identifying objects in an image or video sequence. Humans recognize a multitude of objects in images with little effort, despite the fact that the image of the objects may vary somewhat in different viewpoints, in many different sizes and scales or even when they are translated or rotated. Objects can even be recognized when they are partially obstructed from view. This task is still a challenge for computer vision systems [Pinto et al. 2008]. Many approaches to the task have been implemented over multiple decades. For example, diffusing models to perform image-to-image matching [Thirion 1998], parametric correspondence technique [Barrow 1977] and The Adaptive Least Squares Correlation [Gruen 1985] were presented as a techniques for image matching. Techniques

from [Naphade et al. 1999], [Hampapur et al. 2001] and [Kim et al. 2005] were presented for image sequence matching (video stream).

A related field is visual saliency or saliency detection, “it is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention.” [Laurent, 2007]. Goferman et al. [2012] proposed a new type of saliency/context-aware saliency, which aims at detecting the image regions that represent the scene. In our case, we can exploit the use of eye tracking and its collected data to detect salience in an efficient way since we have fixation points that represents points of interests in a scene.

3. RELATED WORK

As it was mentioned above, many studies were conducted in detecting eye movements before the computer interfaces, as we know them. The studies have been around HCI and usability studies, techniques were presented that can be extended for further eye tracking studies and not just in HCI field. Jacob [1991] presented techniques for eye tracker local calibration which is a procedure of producing a mapping of the eye movements’ measures and wandering in the scene measures. In addition, he presented a techniques for fixation recognition with respect to extracting data from noisy, jittery, error-filled stream and for addressing the problem of "Midas touch” where people look at an item without having the look “mean” something. Jacob and Karn [2003] presented a list of promising eye tracking metrics for data analysis:

- Gaze duration - cumulative duration and average spatial location of a series of consecutive fixations within an area of interest.
- Gaze rate – number of gazes per minute on each area of interest.
- Number of fixation on each area of interest.
- Number of fixation, overall.
- Scan path – sequence of fixations.
- Number of involuntary and number of voluntary fixations (short fixations and long fixations should be defined well in term of millisecond units).

Using handheld devices as a multimedia guidebook in museums has led to improve in the visit experience. Researches have confirmed the hypothesis that a portable computer with an interactive multimedia application has the potential to enhance interpretation and to become a new tool for interpreting museum collections [Evans et al. 2005, Evans et al. 1999, Hsi 2003].

Studies about integration of multimedia guidebooks with eye tracking have already been made in the context of museums and cultural heritage sites. Museum Guide 2.0 [Takumi Toyama et al. 2012] was presented as a framework for delivering multimedia content for museum’s visitors which runs on handheld device and uses the SMI viewX eye tracker device and object recognition techniques. The visitor can hear audio information when detecting an exhibit. A user study was conducted in a laboratory setting, but no real museum was involved. We plan to extend this work by integrating it into a real museum visitors' guide and experiment it is realistic setting.

Brône et al. [2011] have implemented effective new methods for analyzing gaze data collected with eye-tracking device and how to integrate it with object recognition algorithms. They presented a series of arguments why an object-based approach may provide a significant surplus, in terms of analytical precision, precisely they discussed solutions in order to reduce the substantial cost of manual video annotation of gaze behavior, and have developed a series of proof-of-principle case studies in different real world situation, each with their own challenges and requirements. We plan to use their lessons in our study.

Pfeiffer et al. [2014] presented the EyeSee3D method. They combined geometric modelling with inexpensive 3D marker tracking to align virtual proxies with the real-world objects, and this allows classifying fixations on objects of interest automatically while supporting a completely free moving participant. During the analysis of the accuracy of the pose estimation they found that the marker detection may fail from several reasons: First, sometimes the participant looked sideways and there simply was no marker within view. More often, however, swift head movements or extreme position changes were causing these issues. Ohm et al. [2014] tried to find where people look at when navigating in a large scale indoor environment and what objects can assist them to find their ways. They conducted a user study and assessed the visual attractions of objects with an eye tracker. Their findings show that functional landmarks like doors and stairs are most likely to be looked at and named as a landmark. In our case we can use these landmarks as reliable points of interest that can be used for finding the location of the visitor in the museum.

Beugher et al. [2014] presented a novel method for the automatic analysis of mobile eye-tracking data in natural environment and for processing this mobile eye-tracking data by applying object, face and person detection algorithms. The obtained detection results in the object recognition technique were satisfactory for most of the objects. However, a large scale variance results in a lower detection rate (for objects which were looked at both from very far away and from close by.)

Schrammel et al. [2011, 2014] studied attentional behavior of users on the move. They discussed the unique potential and challenges of using eye tracking in mobile settings and demonstrated the ability to use it to study the attention on advertising media in two different situations: within a digital display in public transport and towards logos in a pedestrian shopping street as well as ideas about a general attention model based on eye gaze. Kiefer et al. [2014] also explored the possibility of identifying users’ attention by eye tracking in the setting of tourism – when a tourist gets bored looking at a city panorama – this scenario may be of specific interest for us

as locations or objects that attracted more or less interest may be used to model user's interest and trigger further services/information later on.

Nakano and Ishii (2010) studied the use of eye gaze as an indicator for user engagement, trying also to adapt it to individual users. Engagement may be used as an indicator for interest and the ability to adapt engagement detection to individual users may enable us also to infer interest and build/adapt a user model using this information. Furthermore, Ma et al. [2015] demonstrated an initial ability to extract user models based on eye gaze of users viewing videos. Xu et al. [2008] also used eye gaze to infer user preferences in the content of documents and videos by the users attention as inferred from gaze analysis (number of fixations on word/image).

As we have seen, there is a large body of work about monitoring and analyzing users' eye gaze in general and some also in cultural heritage. Moreover, the appearance of mobile eye trackers opens up new opportunities for research in mobile scenarios. It was also demonstrated in several occasions that eye gaze may be useful in enhancing a user model, as it may enable to identify users' attention (and interests). Considering mobile scenarios, when users also carry smartphones - equipped with various sensors - implicit user modeling can take place by integrating signals from various sensors, including the new sensor of eye-gaze for better modeling the user and offering better personalized services. So far sensors like GPS, compass, accelerometers and voice detectors were used in modeling users' context and interests, (see for instance [Dim & Kuflik. 2014]).

When we mention mobile scenarios, we refer to a large variety of different scenarios – pedestrians' scenario differs from jogging or shopping or cultural heritage (CH) scenario. The tasks are different and users' attention is split differently. The CH domain demonstrates areas where users have long term interests that can be modeled and the model can be used and updated during a museum visit by information collected implicitly from various sensors, including eye-gaze. In this sense, the proposed research extends and aims at generalizing the work of Kardan and Conati [2013].

Still, even though a lot of research effort was invested in monitoring, analyzing and using eye gaze for inferring user interests, so far, little research attention was paid to users gazing behavior "on the go". This scenario poses major challenges as it involves splitting attention between several tasks at the same time – avoiding obstacles, gathering information and paying attention to whatever seems relevant, for many reasons. While users' behavior was monitored and analyzed in various ways in smart environments, using a variety of sensors, this was not done yet about eye gaze.

4. RESEARCH GOAL AND QUESTIONS

Our goal is to examine the potential of integrating the eye tracking technology with a mobile guide for a museum visit and try to answer the question: **How can the use of mobile eye tracker enhance the museum visit experience?** Our focus will be on developing a technique for location awareness based on eye gaze detection and image matching, and integrate it with a mobile museum guide that provide multimedia content to the visitor. For that we will design and develop a system that runs on handheld device and uses Pupil Dev [Kassner et al. 2014] eye tracker for identifying objects of interest and delivering multimedia content to visitor in the museum. Then we will evaluate the system in a user study in a real museum to find out how the use of eye tracker integrated with a multimedia guide can enhance the museum visit experience. In our study, we have to consider different factors and constraints that may affect the performance of the system, such as the real environment lighting conditions which are different from laboratory conditions and can greatly affect the process of object recognition. Another aspect may be the position of the exhibits relative to the eye tracker holder, since the eye tracker device is mounted as this is constrained by the museum layout.

While having many potential benefits, a mobile guide can also have some disadvantages [Lanir et al, 2013]. It may focus the visitor’s attention on the mobile device rather than on the museum artifacts [Grinter et al, 2002]. We will also examine this behavior and try to review whether the use of eye tracker in mobile guide can increase the looking time at the exhibits. In addition, we will try to build a system that runs in various real environments with different factors and have the same constraints such as the light and the position constraints.

5. TOOLS AND METHODS

The study will be a design study [Hevner 2010]. A commercial mobile eye tracker will be integrated into a mobile museum visitors' guide system as a tool for location awareness, interest detection and focus of attention by using computer vision techniques. Our hypothesis is that the use of the eye tracker in mobile guides can enhance the visit experience. The system will be evaluated in user studies, the participants will be students from University of Haifa. The study will be conducted in Hecht museum, which is a small museum, located at the University of Haifa that has both an archeological and art collections. The study will include an orientation about using the eye tracker and the mobile guide, then taking a tour with the eye tracker and handheld device, multimedia content will be delivered by showing information on the screen or by listening to audio by earphones.

Data will be collected as follows: The students will be interviewed and asked about their visit experience, and will be asked to fill questionnaires regarding general questions such as if it is the first time that they have visited the museum, their gender and age, and more.

Visit logs will be collected and analyzed for later use, we can come to conclusions about the exhibit importance and where the visitors tend to look, the positioning of the exhibits, and the time of the visits or explorations.

The study will compare the visit experience when using different systems, we will choose the work of [Wecker et al. 2012] that was conducted in Hecht museum and which uses "light weight" proximity based indoor positioning sensors for location-awareness as a comparison system for examining the user experience.

6. PRELIMINARY RESULTS

It was important to examine the accuracy of eye gaze detection when using the Pupil Dev mobile eye-tracker device. For that, we have conducted several small-scale user studies in order to get an understanding of the performance of the system in realistic setting.

6.1 The Pupil eye tracker

Pupil eye tracker [Kassner et al. 2014] is an accessible, affordable, and extensible open source platform for pervasive eye tracking and gaze-based interaction. It comprises a light-weight eye tracking headset, an open source software framework for mobile eye tracking, as well as a graphical user interface to playback and visualize video and gaze data. Pupil features high-resolution scene and eye cameras for monocular and binocular gaze estimation.

Figure 1. Pupil eye-tracker (http://pupil-labs.com/pupil)

The software and GUI are platform-independent and include state-of-the-art algorithms for real-time pupil detection and tracking, calibration, and accurate gaze estimation. Results of a performance evaluation show that Pupil can provide an average gaze estimation accuracy of 0.6 degree of visual angle (0.08 degree precision) with a processing pipeline latency of only 0.045 seconds.

6.2 User study 1: Look at a grid cells

Five students from the University of Haifa without any visual disabilities participated in this study (average age is 22). They were asked to look at wall-mounted grid from a distance of 2 meters and track a finger (see figure 2. They were standing at a fixed point). On every cell that the finger pointed at, they were asked to look at for approximately 3 seconds. Data was collected for determining the practical measurement accuracy. The results were as follows: on average, fixation detection rate was ~80% (most missed fixations were in the edges/corners – see table 1 for details about misses). In addition, average fixation point error rate, in terms of distance from the center of grids, was approximately 5 cm (exact error rate can be calculated using simple image processing techniques for detecting the green circle and applying mapping transform to the real word – Homography).

#Cell	#Missed
6	5
18	5
19	3
23	5
24	5

Table 1. Experiment details.

Figure 2. Screen capture from user study 1. Finger points at grid where the participant were asked to look at. The green circle is a fixation point given from the eye tracker. The size of each grid is 20x20 cm.

During the study we ran into several practical problems. The Pupil Dev eye tracker device that we are using is not fitted for every person. The device consists of two cameras, the first for delivering the scene and the second directed to the right eye for detecting fixations. In some cases when the device is not fitted correctly, the vision range got smaller and parts of the pupil got out from the capture frame (see figure 3 for example). As a consequence no fixations were detected. Another limitation was that when using the eye tracker with tall persons, where they have to step back from the object which affects the accuracy.

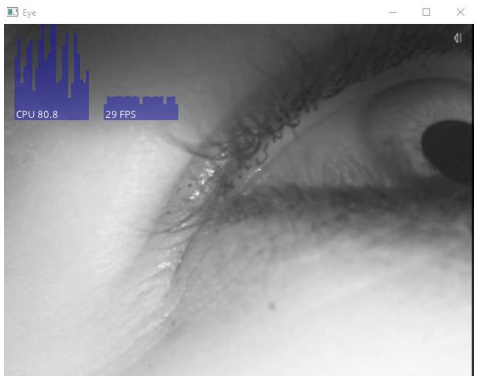


Figure 3. Screen capture from eye camera.

6.3 User study 2: Look at an exhibit

In this user study we have examined the accuracy of the eye tracker in real environment. One participant was asked to look at exhibits in the Hecht museum which located in the campus of University of Haifa. Several exhibits where chosen with different factors and constraints (see figure 4, 5, and 6). The main constraint in this case is the stand distance from the exhibition since the visual range gets larger when we take steps back, and mainly we have to cover all the objects that we are interested in. Collected data are shown in table 2 with regarding to the object height from the floor and the stand distance from the object:

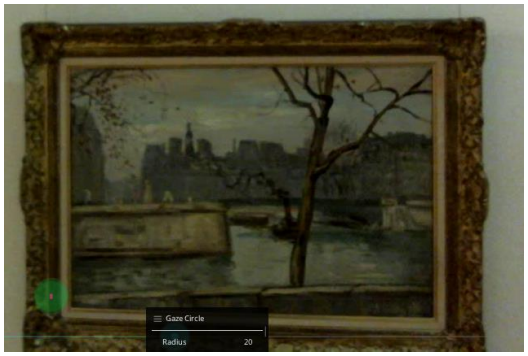


Figure 4. Gallery exhibition

Height (cm)	Stand distance (cm)
160	150
110	230
90	310
40	390

Table 2. Experiment details

It's important to mention that the above heights/distances relation is for visual range and not for fixations detections. Since missed fixations could be as a result of set of constraints and not the distance from the object, thing that we have not examined yet.

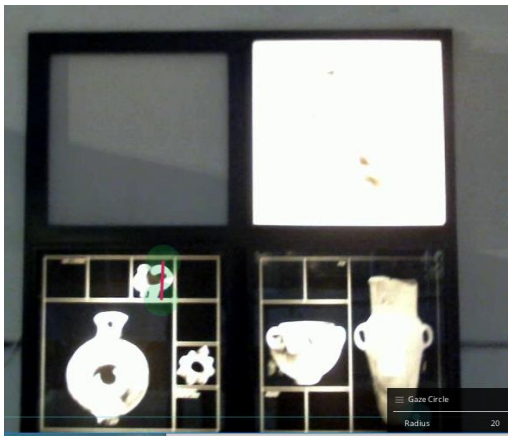


Figure 5. Mounted backlighted images exhibition

Figure 6. Vitrine backlighted exhibition

7. CONCLUSIONS AND FUTURE WORK

This paper presents a work-in-progress that aims at exploring the potential contribution of the mobile eye tracking technology in enhancing the museum visit experience. For that we have done small-scale case studies in order to get an understanding of the performance of the system in realistic setting. We got satisfactory results from using the mobile eye-tracker in a realistic setting. Following the study, we have to take into consideration the limitation of the Pupil-Dev device that we have found when conducting large-scale case user studies.

The next step in the study is to design and build a museum mobile guide that extends the use of mobile eye tracking as a tool for identifying the visitor position and points of interests. We will use the eye-tracker scene camera captures and the collected gaze data to develop a technique for location-awareness. The system will run on tablet, and the participants will listen to audio guide via earphones. Furthermore, knowing exactly where the visitor look in the scene (specific object) will let us deliver personalized information.

Our research will be a supplement to the nowadays mobile museum guide that uses location-awareness technology and techniques that enhances the visit experience. The system can also be extended and used in other venues such as cultural heritage sites and shopping centers/markets after further validation. Future research may address the detection of visits behaviors by analyzing the collected data and logs. Since we will collect and save eye gaze positions, we can build a fixations map and try to detect where the visitors tend to look in the museum in general and in an exhibit in specific.

REFERENCES

[1] Ardissono, L., Kuflik, T., & Petrelli, D. (2012). Personalization in cultural heritage: the road travelled and the one ahead. *User modeling and user-adapted interaction*, *22*(1-2), 73-99.

[2] Barrow, H. G., Tenenbaum, J. M., Bolles, R. C., & Wolf, H. C. (1977). Parametric correspondence and chamfer matching: Two new techniques for image matching (No. TN-153). SRI international, Menlo Park CA, Artificial Intelligence center..

[3] Billings, S. (2009) Upwardly mobile. Mus. Pract., 46, 30–34.

[4] Bitgood, S. (2002). Environmental psychology in museums, zoos, and other exhibition centers. *Handbook of environmental psychology*, 461-480.

[5] Bulling, A., Dachselt, R., Duchowski, A., Jacob, R., Stellmach, S., & Sundstedt, V. (2012, May). Gaze interaction in the post-WIMP world. In CHI'12 Extended Abstracts on Human Factors in Computing Systems, 1221-1224. ACM.

[6] Bulling, A., & Zander, T. O. (2014). Cognition-aware computing. Pervasive Computing, IEEE, 13(3), 80-83.

[7] Cheverst, K., Davies, N., Mitchell, K. and Friday, A. (2000) Experiences of Developing and Deploying a Context-aware Tourist Guide: The GUIDE Project. In Proc. 6th Annu. Int. Conf. Mobile Comput. Netw.,. 20–31. ACM Press, New York.

[8] De Beugher, S., Brône, G., & Goedemé, T. (2014). Automatic analysis of in-the-wild mobile eye-tracking experiments using object, face and person detection. In Proceedings of the international conference on computer vision theory and applications (VISIGRAPP 2014) Vol. 1, 625-633.

[9] Dim, E., & Kuflik, T. (2014). Automatic detection of social behavior of museum visitor pairs. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *4*(4), 17.

[10] E. Veron and M. Levasseur. 1983. Ethographie de l'Exposition. In *Paris, Bibliotheque Publique d'Information*. Centre Georges Pompidou, 91–92.

[11] Evans, J., Sterry, P. : Portable computers & interactive media: A new paradigm for interpreting museum collections. In: Bearman, D., Trant, J. (eds.) Cultural Heritage Informatics 1999: Selected papers from ICHIM 1999, pp. 93–101. Kluwer Academic Publishers, Dordrecht (1999).

[12] Falk, John H., and Lynn D. Dierking. Learning from museums: Visitor experiences and the making of meaning. Altamira Press, 2000.

[13] Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye movements of aircraft pilots during instrument-landing approaches. Aeronautical Engineering Review 9(2), 24–29.

[14] Geert Brône, Bert Oben, Kristof Van Beeck, Toon Goedemé (2011). Towards a more effective method for analyzing mobile eye-tracking data: integrating gaze data with object recognition algorithms. UbiComp '11, Sep 17–Sep 21, 2011, Beijing, China.

[15] Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on,34 (10), 1915-1926.

[16] Grinter, R. E., Aoki, P. M., Szymanski, M. H., Thornton, J. D., Woodruff, A., & Hurst, A. (2002, November). Revisiting the visit: understanding how technology can shape the museum visit. In Proceedings of the 2002 ACM conference on Computer supported cooperative work,146-155. ACM.

[17] Gruen, A. (1985). Adaptive least squares correlation: a powerful image matching technique. South African Journal of Photogrammetry, Remote Sensing and Cartography, 14(3), 175-187.

[18] Hampapur, A., Hyun, K., & Bolle, R. M. (2001, December). Comparison of sequence matching techniques for video copy detection. In Electronic Imaging 2002 (pp. 194-201). International Society for Optics and Photonics.

[19] Hazas, M., Scott, J., & Krumm, J. (2004). Location-aware computing comes of age. Computer, (2), 95-97.

[20] Hendrickson, K., & Ailawadi, K. L. (2014). Six lessons for in-store marketing from six years of mobile eye-tracking research. Shopper Marketing and the Role of In-Store Marketing (Review of Marketing Research, Volume 11) Emerald Group Publishing Limited, 11, 57-74.

[21] Hevner, A., & Chatterjee, S. (2010). Design research in information systems: theory and practice (Vol. 22). Springer Science & Business Media

[22] Kardan, S., & Conati, C. (2013). Comparing and Combining Eye Gaze and Interface Actions for Determining User Learning with an Interactive Simulation. In User Modeling, Adaptation, and Personalization, 215-227. Springer Berlin Heidelberg.

[23] Kassner, M., Patera, W., & Bulling, A. (2014, September). Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. 1151-1160. ACM.

[24] Katy Micha, Daphne Economou (2005). Using Personal Digital Assistants (PDAs) to Enhance the Museum Visit Experience. 10th Panhellenic Conference on Informatics, PCI 2005, Volas, Greece, November 11-13, 2005. Proceedings. 188-198.

[25] Kiefer, P., Giannopoulos, I., Kremer, D., Schlieder, C., & Raubal, M. (2014, March). Starting to get bored: An outdoor eye tracking study of tourists exploring a city panorama. In Proceedings of the Symposium on Eye Tracking Research and Applications (pp. 315-318). ACM.

[26] Kim, C., & Vasudev, B. (2005). Spatiotemporal sequence matching for efficient video copy detection. Circuits and Systems for Video Technology, IEEE Transactions on, 15(1), 127-132.

[27] Kuflik, T., Lanir, J., Dim, E., Wecker, A., Corra, M., Zancanaro, M., & Stock, O. (2012, November). Indoor positioning in cultural heritage: Challenges and a solution. In Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of. 1-5. IEEE.

[28] Lanir, J., Kuflik, T., Dim, E., Wecker, A. J., & Stock, O. (2013). The influence of a location-aware mobile guide on museum visitors' behavior. Interacting with Computers, 25(6), 443-460.

[29] Laurent Itti (2007) Visual salience. Scholarpedia, 2(9):3327.

[30] Leinhardt, Gaea, and Karen Knutson. Listening in on museum conversations. Rowman Altamira, 2004.
[31] Ma, K. T., Xu, Q., Li, L., Sim, T., Kankanhalli, M., & Lim, R. (2015). Eye-2-I: Eye-tracking for just-in-time implicit user profiling. arXiv preprint arXiv:1507.04441.
[32] McManus, Paulette. "Towards understanding the needs of museum visitors."The manual of museum planning, London, HMSO (1991).
[33] Nakano, Y. I., & Ishii, R. (2010, February). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In Proceedings of the 15th international conference on Intelligent user interfaces. 139-148. ACM.
[34] Naphade, M. R., Yeung, M. M., & Yeo, B. L. (1999, December). Novel scheme for fast and efficent video sequence matching using compact signatures. InElectronic Imaging. 564-572. International Society for Optics and Photonics.
[35] O. Stock, M. Zancanaro, F. Pianesi, D. Tomasini, and C. Rocchi. 2009. Formative evaluation of a tabletop Display meant to orient casual conversation. Journal of Knowledge, Technology and Policy 22, 1, 17–23.
[36] Ohm, C., Müller, M., Ludwig, B., & Bienk, S. (2014). Where is the Landmark? Eye Tracking Studies in Large-Scale Indoor Environments.
[37] Packer, Jan, and Roy Ballantyne. "Solitary vs. shared: Exploring the social dimension of museum learning." Curator: The Museum Journal 48.2 (2005): 177-192.
[38] Pinto, Nicolas, David D. Cox, and James J. DiCarlo. "Why is real-world visual object recognition hard?." (2008): e27.
[39] Pfeiffer, T., & Renner, P. (2014, March). Eyesee3d: A low-cost approach for analyzing mobile 3d eye tracking data using computer vision and augmented reality technology. In Proceedings of the Symposium on Eye Tracking Research and Applications. 369-376. ACM.
[40] ROBERT J. K. JACOB (1991). The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look At is What You Get, ACM Transactions on Information Systems, Vol. 9, No 3, April 1991, 152-169.
[41] Robert J. K. Jacob, Keith S. Karn (2003). Eye Tracking in Human–Computer Interaction and Usability Research: Ready to Deliver the Promises, Elsevier Science BV.
[42] Schrammel, J., Mattheiss, E., Döbelt, S., Paletta, L., Almer, A., & Tscheligi, M. (2011). Attentional behavior of users on the move towards pervasive advertising media. In Pervasive Advertising (pp. 287-307). Springer London.
[43] Schrammel, J., Regal, G., & Tscheligi, M. (2014, April). Attention approximation of mobile users towards their environment. In CHI'14 Extended Abstracts on Human Factors in Computing Systems (pp. 1723-1728). ACM
[44] S.Hsi (2003). A study of user experiences mediated by nomadic web content in a museum. The Exploratorium, 3601 Lyon Street, San Francisco, CA 94123
[45] Stephens, S. (2010). The growth of mobile apps. Mus. Pract.
[46] Takumi Toyama, Thomas Kieninger, Faisal Shafait, Andreas Dengel. Gaze guided object recognition using a head-mounted eye tracker. ETRA '12 Proceedings of the Symposium on Eye Tracking Research and Applications, 91-98, ACM New York, NY, USA 2012
[47] Thirion, J. P. (1998). Image matching as a diffusion process: an analogy with Maxwell's demons. Medical image analysis, 2(3), 243-260.
[48] Xu, S., Jiang, H., & Lau, F. (2008, October). Personalized online document, image and video recommendation via commodity eye-tracking. In Proceedings of the 2008 ACM conference on Recommender systems. 83-90. ACM.
[49] Yalowitz, S.S. and Bronnenkant, K. (2009) Timing and tracking: unlocking visitor behavior. Visit. Stud., 12, 47–64.
[50] Zhang, Z., Deriche, R., Faugeras, O., & Luong, Q. T. (1995). A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Artificial intelligence, 78(1), 87-119.

CANE: A Wearable Computer-Assisted Navigation Engine for the Visually Impaired

Seth Polsley
Sketch Recognition Lab
Texas A&M University
spolsley@tamu.edu

Vijay Rajanna
Sketch Recognition Lab
Texas A&M University
vijay.drajanna@gmail.com

Larry Powell
Sketch Recognition Lab
Texas A&M University
larry.powell@tamu.edu

Kodi Tapie
Sketch Recognition Lab
Texas A&M University
kstapie@tamu.edu

Tracy Hammond
Sketch Recognition Lab
Texas A&M University
hammond@tamu.edu

ABSTRACT
Navigating unfamiliar environments can be difficult for the visually impaired, so many assistive technologies have been developed to augment these users' spatial awareness. Existing technologies are limited in their adoption because of various reasons like size, cost, and reduction of situational awareness. In this paper, we present **CANE**: "Computer Assisted Navigation Engine," a low cost, wearable, and haptic-assisted navigation system for the visually impaired. CANE is a "smart belt," providing feedback through vibration units lining the inside of the belt so that it does not interfere with the user's other senses. CANE was evaluated by both visually impaired and sighted users who simulated visual impairment using blindfolds, and the feedback shows that it improved their spatial awareness allowing the users to successfully navigate the course without any additional aids. CANE as a comprehensive navigation assistant has high potential for wide adoption because it is inexpensive, reliable, convenient, and compact.

Author Keywords
Visually impaired; Wearable; Haptic assisted; Navigation system; Ultrasonic sensors; Non-intrusive; Ubiquitous; Internet of Things

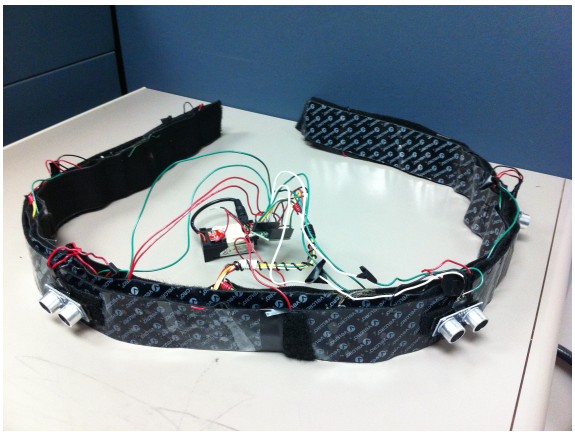
ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—*Haptics I/O*

IUI 2016 Workshop: A joint Workshop on Smart Connected and Wearable Things, March 10th, 2016, Sonoma, CA, USA

urn:nbn:de:tuda-tuprints-54208

INTRODUCTION
Most people rely heavily on their sight when navigating unfamiliar environments. While vision is an important sense in many daily tasks, it becomes crucial to the capability to explore and wander new places freely. Even with other senses to help cope with the loss of sight, the visually impaired can still be limited in their ability to explore new environments. Studies have shown that only about one quarter of working age blind individuals are employed [20] and nearly every blind individual can struggle with day-to-day tasks at some point. It is important for the visually impaired to be able to explore in order to be immersed in ordinary life, but it is easy to imagine how difficult wandering new places can be. "My biggest fear is being alone in a huge parking lot with nothing around me and no sounds," says one blind man from our user study about exploring outside. Because of a combination of healthcare costs, concerns over quality of life, and the fact that the number of visually impaired individuals increases yearly [11], there is a strong motivation to help solve the problems faced by these individuals.

Researchers have spent years trying to find better ways to accommodate the visually impaired to ensure that they can enjoy the same opportunities and capabilities as normal-sighted individuals; independent mobility is one such valuable capability. There are several traditional solutions outlined by Strumillo [15] that have long been available, but these methods have their limitations. Guide dogs have been used for many years as a means of helping the visually impaired, but in addition to needing care like any other dog, their training is time consuming and expensive. A human caregiver can help alleviate many issues faced by the visually impaired, but this solution inherently limits independence. White canes are a common alternative that allow their users to reclaim some independence, but despite the low cost and efficiency in detecting obstacles, there are still restrictions on how much canes can detect, especially at levels above the waist. Unsurprisingly, researchers have turned to technology in recent years as a possible means of navigation for the visually impaired that is inexpensive, convenient, and reliable.



a

Figure 1: The complete CANE system with the sensors on the outside edge and the vibration units hidden inside.

This work introduces the Computer-Assisted Navigation Engine (CANE) for assisting visually impaired individuals in the form of a wearable "smart belt." As with any such system, CANE is intended to promote increased mobility and independence, but it was built around several design principles that provide advantages over other blind navigation technologies. First and foremost is the advantage of it being a wearable device that can interact with other mobile devices over wireless communication. CANE uses on-board ultrasonic sensors and vibration motors to give users vibrotactile-based directions in an indoor environment, and the system can be further extended to integrate with the Internet of Things (IoT) for enhanced navigational support. The built-in sensors can detect the proximity of objects to the user, as well as their location within the sensor's field of view. This information is used to provide real-time vibrotactile feedback for guiding the user. Because the sensors are spread around the front, CANE reduces the need for sweeping motions, such as those made with physical canes or many other sensor-equipped feedback systems. Furthermore, not only is it designed to be lightweight and inexpensive, the fact that it can be worn as a normal belt makes it less conspicuous than larger, more complex systems. This is also aided by its simple design that excludes headsets, helmets, or other highly noticeable components. Because the visually impaired rely strongly on their other senses, such as sound, CANE does not use any audio feedback or cover their face, which could be distracting. By using the sense of touch, CANE behaves much more like traditional navigational tools like physical canes or guide dogs. A pictorial depiction of CANE is shown in Figure 1. In order to understand the robustness and potential impact of CANE, it was tested on normal-sighted users first and then on visually impaired users. The sighted users were asked to navigate blindfolded in an unfamiliar environment while being assisted by CANE. During post-study interviews a majority of these users shared that, although they were apprehensive at the beginning of the study, they adapted quickly to using CANE. Further, CANE was tested by two visually impaired individuals. Experiences shared following testing with both the normal-sighted and visually impaired users showed that the feedback from CANE was reliable and intuitive, although additional sensors may be beneficial. In particular, users liked CANE's affordability, wearability as a belt, and small form factor.

The remaining sections are organized as follows. First, we will discuss the related works in the area of assistive technologies for the visually impaired and how they differ from CANE. After the discussion of related works we describe the system design and implementation. The experimental design section then explains the procedures followed during the user studies. We subsequently present the results of our user studies and discuss the key findings. Finally, we conclude with a brief discussion of the areas of improvement for CANE, and the future of low cost, wearable, inconspicuous, assistive technology for the visually impaired.

PRIOR WORK

Many visual impairment navigation technologies have been developed that combine wearables with sensors, global positioning, and some means of feedback for guiding users. Unfortunately, a technological solution has yet to gain widespread adoption, which is perhaps due to issues with many current solutions.

Cost and size are significant barriers for some technologies. Willis and Helal proposed a grid of RFID chips be created to pair with their wearable system for navigation, but the RFID chips must be installed everywhere the system is intended to be used [20]. This carries a large upfront cost and necessary infrastructure before users would begin to benefit in unfamiliar terrain. Ran et al., proposed Drishti, a system combining ultrasound and GPS for both indoor and outdoor navigation [11]. While an ultrasound-based wearable is less expensive than requiring an external grid, Drishti weighs around 8 pounds and includes a waist strap and headset for voice control and audio feedback, making it very noticeable. Some visually impaired users may be hesitant to wear such a device due to its size and weight; they may also feel more comfortable with smaller devices that are less conspicuous. It has been shown that blindness can carry with it physical and mental health implications, and it can have an impact on an individual's self-esteem [17]. Related to the size of a navigation system is the number of components. Dakopoulos and Bourbakis presented four major findings in his survey of visually impaired navigation and obstacle detection systems, and one of them is that the system should be hands-free [5]. He argues that users will always feel more comfortable if they have the option to hold a traditional cane alongside the system. Tools like Borenstein's second iteration of the Navbelt included a cane as part of the navigation system, restricting the user's choice [14]. The limitations of these works reflect that guidance systems should aim to be subtle, small, and lightweight.

The feedback system is also very important to consider when looking at these technologies. Two of Dakopoulos' and Bourbakis' other findings both relate to the feedback system—it should be simple and not obstruct hearing [5]. Simplicity is an important concept in every human-computer interaction system. Keeping the ears free is a less obvious requirement,

and many existing systems ignore this design recommendation. However, research has shown that the blind come to rely more heavily on their hearing than normal-sighted individuals, with parts of their visual cortex even being re-purposed for auditory processing [16]. Given the importance of sound in their interactions with the environment, a navigation system that provides auditory feedback may actually be an impediment. One of the earliest ultrasonic-based navigation systems, Navbelt from the University of Michigan, required the user to wear a set of headphones to receive the feedback from the sensors [13]. The successor system also could only provide feedback via headphones [14]. NOPPA, developed by Ari and Sami, is a mobile guidance system that pairs obstacle detection and GPS to direct users [18]; unfortunately, its only means of interfacing is done through a speaker and microphone. Even Drishti included a headset [11].

Some systems have been built that use haptic feedback rather than audio. Haptics is a promising approach that solves both of Dakopoulos' and Bourbakis' concerns about the interface. First, the user's hearing is unimpeded, and hence his or her situational awareness remains fully intact. Second, touch can be a very intuitive means of communication, as seen in the widespread usage of touch-enabled devices today. Colwell, et al., created a haptic device for exploring virtual worlds and three-dimensional objects through touch, rather than navigating real-world environments [2]. Similarly, Lahav, et al., used haptics to help blind users build mental maps of new environments, but the technology does not provide real-time feedback to new environments [8]. Many other works have examined the usage of haptics as a means of increasing mobility for the visually impaired and encouraging the development of mental maps, but much of this work focuses around virtual environments [3, 7, 12]. One of the earliest wearable haptic devices was developed by Ertan, et al. They created a vest that provided a physical map of the environment through a vibrotactile array on the user's back [6]. Unfortunately, this system required ceiling-mounted infrared transceivers to track the user's location and a stored map of the environment in the vest's memory. A similar work was developed by Dakopoulous, et al., in [4] where they used a similar vibrotactile array but received the environment's mapping from a camera-based system. The algorithm was reliable, but users had difficulty interpreting their location relative to obstacles. Researchers at the University of Toronto built a head-mounted vibrotactile system that uses a Kinect to interpret the user's surroundings [9]. While the collision avoidance capabilities are promising, a headset-based solution interferes with hearing and is conspicuous. HALO (Haptic Alerts for Low-hanging Obstacles), an incremental work, integrates ultrasonic sensors and vibrators with a traditional white cane to warn of overhead obstacles [19]. There exist commercial products that perform a similar function, like the SmartCane™ [1]. One of the most similar systems to CANE is the wearable vest developed by Cardin, et al. Like CANE, it features four ultrasonic sensors and eight vibrotactile motors for feedback, but they are placed on the chest, limiting obstacle detection strictly to the horizontal plane at chest level [1]. Another wearable vest using haptic feedback was constructed by Prasad, et al., that also used chest-level sensors for obstacle detection and paired with GPS to provide directions [10].

Unsurprisingly, there has been a lot of research using ultrasonic sensors and audio or haptic feedback as they apply to navigation solutions for the visually impaired. Most haptic-equipped navigation systems have been developed only in recent years, but none of these have been able to present a low-cost, standalone solution that is also inconspicuous and non-restrictive on the user's situational awareness. While it cannot address each concern as well as every system, CANE is designed to provide a balance of all of these aspects, making it a powerful option.

SYSTEM ARCHITECTURE

CANE is a "smart belt" – a wearable, multi-agent, haptics-based assisted navigation system for the visually impaired. CANE uses state of the art technologies like *sonar sensors*, *vibrotactile motors*, and a microcontroller to create a robust, intelligent system that enhances spatial awareness of the visually impaired. One important aspect of CANE that bears mention is its low-cost. The whole system was built at a cost of $59 comprising of four main electronic components: a) Vibrotactile motors: $16, b) Ultrasonic sensors: $8, c) Teensy microcontroller: $30, and d) Wires and solder: $5. The system consists of three modules working together: 1) an input array of four ultrasonic sensors, 2) a feedback array of eight vibrotactile motor, and 3) a central control system. The outer module is an array of ultrasonic sensors. These sensors face outward from the front of the belt and are constantly receiving feedback from the environment. This feedback includes information like which sensor is active, for direction purposes, and the distance to an obstacle. The input from the sensor array is used to drive the inner module, the vibration units. There are a total of eight equidistant vibration units lining the inner layer of the belt. Vibrators are activated based on the information provided by the outer sensors, and their intensity level changes according to the proximity of the obstacle. The most important module is the central control system that connects the input to the output, which is implemented on a Teensy++ 2.0 board running software which parses the data provided by the sensor array. It uses the distance measure from a given sensor to activate the associated vibrotactile motors with an appropriate intensity. This module handles most of the system logic and may be extended for communication with other devices for enhanced navigation modes through IoT. A pictorial depiction of the working model is shown in Figure 2.

SYSTEM IMPLEMENTATION

Input Layer of Ultrasonic Sensors

A total of four HC-SR04 ultrasonic sensors[2] are included on CANE to detect obstacles. In early iterations, one pair of sensors was placed outward on the front while the other pair was placed on the back of the belt. Following initial testing, later implementations of CANE moved the sensors more frontward, since the back sensors served little purpose in their

[1] smartcane.saksham.org/overview/

[2] micropik.com/PDF/HCSR04.pdf

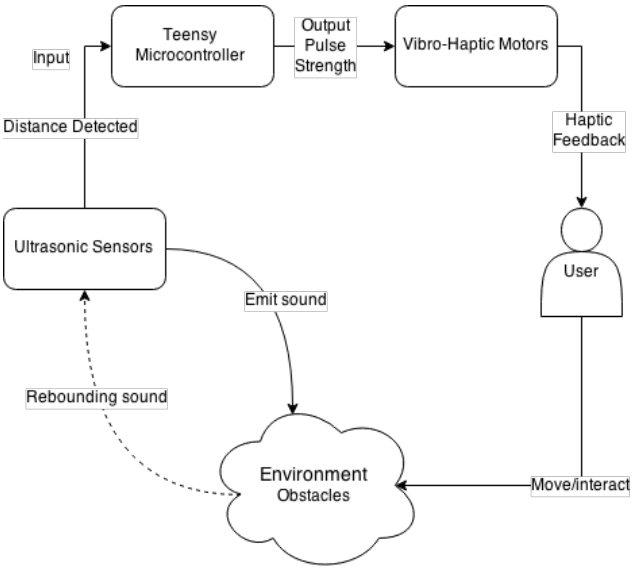


Figure 2: The working model for CANE

original location. For each pair, the sensors are intended to angle away from the center at approximately a 30-45 degree angle on each side. The sensors connect to the belt using velcro in specific regions on the surface of the belt. One consequence of this design is that the sensors' angles will vary based on an individual's waist size. The HC-SR04 sensors activate localized vibrotactile motors on the inside of the belt. These sensors provide a distance measure from their echo port, which the software uses to determine the intensity of the haptic feedback. A braided wiring harness connects the sensors to the inner layer of the belt in order to reduce clutter.

Feedback Layer of Vibrotactile Motors

The feedback layer is an array of eight LilyPad Vibe Boards[3]. Input from a single sonar sensor controls the activation and deactivation of two vibration motors. The vibration motors are arranged on the inward surface of CANE that makes contact with the wearer's body. Locations of the vibration motors correspond to the locations of the sonar sensors; this arrangement provides an accurate estimate of the direction and orientation of the obstacle.

Central Control System

The central control system is implemented using a Teensy++ 2.0 microcontroller; it establishes a common platform for all the components to interact. Unlike the ultrasonic sensors and the vibration units, the microcontroller is not embedded inside the belt. Current iterations of CANE place the Teensy++ 2.0 on top of the center of the belt, allowing easy access to the system's power and pin reconfigurability. The microcontroller constantly receives trigger input signals from all four ultrasonic sensors, which it then maps to specific vibrators. There are a total of three intensity levels, that are dynamically calculated based on the distance of the obstacle from

[3] sparkfun.com/products/11008

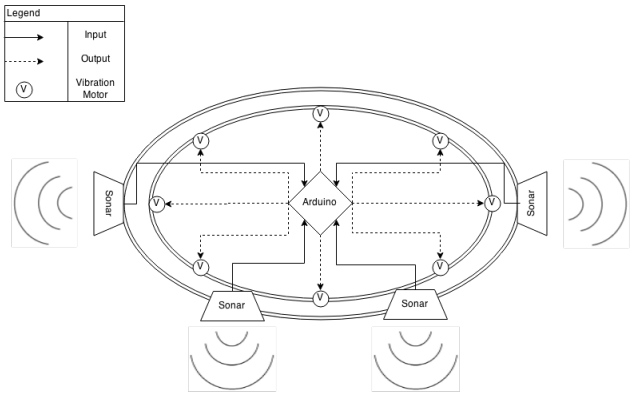


Figure 3: A diagram view of CANE's implementation

the wearer. Based on values determined empirically, the vibration motors only become active when an obstacle is within 80 cm, which corresponds to the lowest level of vibration intensity. After the mid-range level, which begins at 65 cm, the motors will provide maximum vibration for obstacles within 50 cm. A pictorial depiction of the system implementation is shown in Figure 3.

EXPERIMENT DESIGN

CANE's evaluation had two goals to explore: 1) test the validity of the proposed design principles and 2) cultivate a deeper understanding of the traits necessary for an assistive technology to address blind navigation. Hence, the system evaluation was divided into two phases. First, a pool of sighted participants tested the system and provided their feedback. This study was conducted under laboratory settings that simulated real-world scenarios in a controlled environment. A group of 25 university students participated in the study; that total includes 22 males and 3 females aged between 18 to 25. One of these participants was legally blind. Users were tasked with navigating a single lap around a square hallway while blindfolded. Each side of the hallway was approximately 10 meters in length. After walking straight down a short walkway, users had to make a right turn into the hallway, navigate obstacles, and make appropriate left turns to traverse the path in a counter-clockwise fashion. Before testing began, users were instructed about the navigation task, but no information was provided regarding functionality of the system. In doing so, we aimed to gauge the learning curve and users' spatial awareness based on the intensity and relative positioning of haptic feedback. Using the *think-aloud* strategy, we constantly sought users' feedback as they navigated the hallway. Upon completion of the task, each user was briefed on the design principles and the system's functionality. During the second phase, the system was tested by one completely blind participant, a male aged 60. He had lost his sight in an accident at 31 years of age. The blind participant tested the system at his residence and validated the feedback provided by the normal-sighted participants. He was also able to provide more key insights by evaluating the system from a blind user's perspective, as well as providing an experience-based

comparison to some of the other navigation technologies he had used.

RESULTS AND DISCUSSION

Sighted User Testing

While having sighted users wear a blindfold is not an accurate simulation in every regard, the purpose of this first round of testing was primarily to find flaws and identify issues that could be fixed in the subsequent versions of CANE. Hence, these tests were conducted over several iterations of the device. A significant amount of feedback was received as the tests included 25 users with distinct statures and physiological conditions. Users quickly learned to interpret the vibrotactile feedback and navigated confidently, while successfully moving away from obstacles. The short learning curve supports our reasoning that the vibrotactile approach is intuitive. We also found that since CANE is worn around the waist, the varying heights of the users limit its ability to detect objects at knee height. An attempt was made to address this issue by pointing the sensors at a partial downward angle. However, a downward orientation raises two issues: 1) limited range of sensors, and 2) a larger activation distance when compared to the outward orientation. We believe that including sensors with varying levels of sensitivity could alleviate this problem.

Some users had a related concern regarding the placement of the vibration units. As the sensors were placed equidistantly inside the belt, some users with larger waists found the placement sparse. This issue was countered with minor adjustments of vibrator positions, but more vibrators could be added to reduce the likelihood of large gaps. As we expected, most users were able to navigate the complete course using CANE; further, they approved of its wearability and small form factor. It was unsurprising that sighted users, being blindfolded, moved slowly while they completely relied on CANE. An interesting revelation is that, despite their ability to avoid obstacles, they still preferred walking along one side of the hallway as they used the wall as their reference and stayed just close enough to receive constant feedback. Another useful result from these studies showed that sensors on the back side of the belt added no value, and in the later versions, these sensors were moved frontward. However, when sensors are placed directly on the side of the belt, they can meet interference from users' arms, implying that they could be moved further frontward.

Blind User Testing

A user who is blind evaluated CANE by using it as a navigational aid in a familiar environment. This test also included an interview component, both before and after using CANE. The initial discussion centered around existing navigational tools for the visually impaired, in terms of their availability, usability, and limitations. The user has both a guide dog and a cane, which constitute his primary means of navigation. He had formerly used two ultrasonic devices, both of which were similar in design to the SmartCane™, but one placed the sensor lower on the shaft while the other replaced the shaft with a longer-range, downward-pointing sensor. Both had to be used like a normal cane, requiring a constant sweeping pattern, so the additional cost of the ultrasonic sensors did not add value to the user. However, he did appreciate their ability to detect curbs and stairway edges like a typical cane. "Gravity never fails!" is the motto he repeated several times, and it is one of the first lessons taught when learning to navigate blind. Confidently being able to detect drops in the floor was his foremost concern, more so than detecting obstacles.

Following a brief initial interview, we asked the user to test CANE by walking around his home, a well-known environment, only by using the feedback provided by the belt. This test was run at his home primarily to ensure his safety, and it allowed him to validate CANE's feedback with his own spatial knowledge of the environment. He seemed very comfortable with the existing design and wearability of CANE, but he found the waist-level feedback to be too high on the body under some circumstances. For instance, downward pointing sensors would give more assurance about the ground, and one additional complication was the level at which he held his hands. When walking without his guide dog or cane, he would hold his arms in front of him to give early detection of obstacles. Unfortunately, this could lead to false positives for the presence of an obstacle. The angle of approach could also be a concern; because the sensors are angled to detect side obstacles, they can miss certain objects like corners.

An interesting behavior emerged during testing in which the user performed slight sweeping motions of the upper body. The intended goal of the sensor placement was to remove the sweeping pattern of canes, but the user still rotated his body to build a small mental map of the objects around him. He said this was important because it gave a better sense of his environment in a manner which he controlled. This leads to a larger discussion on the importance of the sweeping motion to a visually impaired user. CANE can reduce this need by providing more sensors, but sweeping may always be a component of the navigation process for the visually impaired.

A small closing discussion followed the test. The user mentioned that his main concern using CANE was that the range of the sensors seemed too limiting. "You would not want to move very fast," he commented. This limitation appeared during the first round of user testing as well, suggesting that it is not just tied to the user's comfort walking without sight. Longer-range sensors will be important in providing earlier warnings of obstacles so that the wearer can move around more confidently. The user liked the varying levels of intensity that CANE provides, and this intensity scaling model fits with distinguishing between farther and nearer objects. Finally, he noted that CANE could help build user assurance by giving a constant pulsing feedback to signify it is still operational, similar to how cane users tap walls every few steps to ensure safety. This validates the earlier finding that sighted users stayed close enough to a wall to receive constant feedback as they moved. Overall, CANE was well-received by the test users. We found that providing more sensors with longer range and adjustable orientations would make it more applicable to many real-world environments. The user results did indicate that CANE has an excellent form factor, is very lightweight, and gives useful feedback to supplement spatial awareness.

(a) (b)

Figure 4: Images from the visually impaired user study showing a user wearing CANE from the (a) front and (b) side.

FUTURE WORK

There are many improvements to be made in CANE. First, as validated in testing, it will be important to redesign the sensor layout on the belt. The current layout can miss corners and obstacles depending on the angle and height relative to the person. By adding more sensors and making the locations or angles more adjustable, users will be able to move better in greatly varied environments. Also, the distance of the sensors can be limiting to the user's speed; using longer-range ultrasonic sensors may allow the user to move more quickly by providing more timely feedback. Second, while it is flexible and wearable, to make the belt easier to use, the system would benefit from stitching the wiring into the lining. Once the sensor configuration is finalized, the microcontroller can also be placed inside the belt and the battery compartment moved to a more convenient place. We also hope to build a smartphone application to pair with the microcontroller to make it easier to configure and extend with other navigational modes over the Internet. Finally, more studies will be conducted with visually impaired users so that CANE can be designed to suit more people's needs. In order to evaluate CANE's performance in varying environments, these studies will be run with different scenarios and obstacle types, including new measures like timing.

CONCLUSION

Navigating unfamiliar environments without sight can be difficult. While many tools and technologies exist that try to address this issue, many of them are restricted by issues like expense, size, or usability. In this work we have presented CANE: "Computer Assisted Navigation Engine," a low-cost, wearable, non-intrusive, and haptic-assisted navigation system for the visually impaired. CANE is a "smart belt" that is fitted with an array of four ultrasonic sensors on its surface to identify obstacles, their relative locations, and an approximate estimate of the distance to the user. This localized knowledge of the environment is delivered to the user through eight vibrotactile motors along the inside surface of the belt. The locations of the sensors and vibrators align so that the haptic feedback naturally augments the user's spatial awareness. The system was evaluated in two phases, first by a pool of 25 sighted participants simulating blindness and second by a blind participant. Both studies yielded many interesting findings that will be used to direct future development. Among these is that users value constant feedback to assure them the system is working and that they are on the right path. Additionally, greater user control in placement, range, and direction of the sensors would help CANE appeal to more users in many environments. CANE was found to be a useful system that could aid in navigation for the visually impaired both inexpensively and inconspicuously. It balances a combination of many factors that other such projects have failed to consider. When paired with the current IoT trends toward smaller, faster, and wearable devices, these results show that technological tools may soon replace more traditional solutions and make navigating the world much easier for the visually impaired.

ACKNOWLEDGEMENTS

We acknowledge Matt Harper, Grant Hendley, and Jeff Harrison for their initial work on CANE. We are very grateful to Dr. Nancy Leslie for her guidance during the course of this work. Additionally, Cassandra Odoula and David Turner were instrumental in this project. We also thank the members of the Sketch Recognition Lab, the IAP members, and the Computer Science and Engineering Department at Texas A&M University for their feedback and support.

REFERENCES

1. Cardin, S., Thalmann, D., and Vexo, F. A wearable system for mobility improvement of visually impaired people. *The Visual Computer 23*, 2 (2007), 109–118.

2. Colwell, C., Petrie, H., Kornbrot, D., Hardwick, A., and Furner, S. Haptic virtual reality for blind computer users. In *Proceedings of the third international ACM conference on Assistive technologies*, ACM (1998), 92–99.

3. Crossan, A., and Brewster, S. Two-handed navigation in a haptic virtual environment. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, ACM (2006), 676–681.

4. Dakopoulos, D., Boddhu, S., and Bourbakis, N. A 2d vibration array as an assistive device for visually impaired. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, IEEE (2007), 930–937.

5. Dakopoulos, D., and Bourbakis, N. Wearable obstacle avoidance electronic travel aids for blind: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 40*, 1 (2010), 25–35.

6. Ertan, S., Lee, C., Willets, A., Tan, H., and Pentland, A. A wearable haptic navigation guidance system. In *Wearable Computers, 1998. Digest of Papers. Second International Symposium on*, IEEE (1998), 164–165.

7. Lahav, O., and Mioduser, D. A blind person's cognitive mapping of new spaces using a haptic virtual environment. *Journal of Research in Special Educational Needs 3*, 3 (2003), 172–177.

8. Lahav, O., and Mioduser, D. Haptic-feedback support for cognitive mapping of unknown spaces by people who are blind. *International Journal of Human-Computer Studies 66*, 1 (2008), 23–35.

9. Mann, S., Huang, J., Janzen, R., Lo, R., Rampersad, V., Chen, A., and Doha, T. Blind navigation with a wearable range camera and vibrotactile helmet. In *Proceedings of the 19th ACM international conference on Multimedia*, ACM (2011), 1325–1328.

10. Prasad, M., Taele, P., Olubeko, A., and Hammond, T. Haptigo: A navigational 'tap on the shoulder'. In *Haptics Symposium (HAPTICS), 2014 IEEE* (Feb 2014), 339–345.

11. Ran, L., Helal, S., and Moore, S. Drishti: an integrated indoor/outdoor blind navigation system and service. In *Pervasive Computing and Communications, 2004. PerCom 2004. Proceedings of the Second IEEE Annual Conference on*, IEEE (2004), 23–30.

12. Sánchez, J., and Tadres, A. Audio and haptic based virtual environments for orientation and mobility in people who are blind. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, ACM (2010), 237–238.

13. Shoval, S., Borenstein, J., and Koren, Y. Auditory guidance with the navbelt-a computerized travel aid for the blind. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 28*, 3 (1998), 459–467.

14. Shoval, S., Ulrich, I., and Borenstein, J. Navbelt and the guide-cane [obstacle-avoidance systems for the blind and visually impaired]. *Robotics & Automation Magazine, IEEE 10*, 1 (2003), 9–20.

15. Strumillo, P. Electronic interfaces aiding the visually impaired in environmental access, mobility and navigation. In *Human System Interactions (HSI), 2010 3rd Conference on* (May 2010), 17–24.

16. Théoret, H., Merabet, L., and Pascual-Leone, A. Behavioral and neuroplastic changes in the blind: evidence for functionally relevant cross-modal interactions. *Journal of Physiology-Paris 98*, 1 (2004), 221–233.

17. Tuttle, D. W., and Tuttle, N. R. *Self-esteem and adjusting with blindness: The process of responding to life's demands*. Charles C Thomas Publisher, 2004.

18. Virtanen, A. Navigation and guidance system for the blind. *Proceedings of Interactive Future and Man 1* (2003).

19. Wang, Y., and Kuchenbecker, K. J. Halo: Haptic alerts for low-hanging obstacles in white cane navigation. In *Haptics Symposium (HAPTICS), 2012 IEEE*, IEEE (2012), 527–532.

20. Willis, S., and Helal, S. Rfid information grid for blind navigation and wayfinding. In *ISWC*, vol. 5 (2005), 34–37.

Enhanced sliding window approach for the inertial-based activity recognition

Sergey Zeltyn
IBM Research
Mount Carmel, Haifa
sergeyz@il.ibm.com

Lior Limonad
IBM Research
Mount Carmel, Haifa
liorli@il.ibm.com

Alexander Zadorojniy
IBM Research
Mount Carmel, Haifa
zalex@il.ibm.com

ABSTRACT
We address real-time recognition of human activities using inertial devices, such as accelerometers and gyroscopes. A conventional sliding-window approach to this type of activity recognition is generalized by introducing a new family of enhanced recognition rules. We base activity recognition on data analysis from several consecutive sliding windows rather than from a single window, and determine the activity class based on detection over a threshold number of intervals (e.g., three-out-of-five). After evaluating these enhanced rules on an extensive benchmark data set [4], we found they perform better than the conventional approach for a pool of eight special cases.

Our enhanced rules, together with clearly specified parameters, potentially yield a formal framework for effective activity recognition, given certain criteria for optimality and performance.

Author Keywords
Wearable computing; inertial sensing; sliding windows; segmentation; activity recognition.

ACM Classification Keywords

H.1.2. Human information processing.

ACTIVITY RECOGNITION: CHALLENGES WITH THE SLIDING WINDOW APPROACH

Real-time recognition of human activities is a popular and very broad research area. It gives rise to important applications in various fields, including: industrial safety, elderly care, public security, safe driving, and others.

Different technological means, such as motion sensors or video cameras, can be employed to recognize activity. In this paper, we discuss activity recognition based on body-worn inertial sensors. The most common examples of inertial sensors are the 3-axis accelerometer and 3-axis gyroscope [5]. The former measures acceleration and the latter measures angular velocity. These sensors are used to detect typical body movement patterns for different activities, gestures, and events. Such devices are relatively low-cost and highly useful for many application purposes such as: fall detection to aid with industry safety and care of the elderly [8]; gesture recognition to enable hands-free operations by field workers [9]; fatigue detection for industry and service employees, where movement patterns may indicate the fatigue level.

IUI 2016 Workshop: A joint Workshop on Smart Connected and Wearable Things, March 10th, 2016, Sonoma, CA, USA

urn:nbn:de:tuda-tuprints-54208

Inertial devices can be used either alone or together with physiological sensors that provide additional measurements such as heart rate, body temperature, galvanic skin response, and other metrics. Video footage is another potential source of information on human activities. In this work we restrict the focus to recognition based solely on inertial devices.

Inertia-sensing is a promising technology, but the real question is how to make it practical for business purposes. Currently, its success is restricted to several domains, primarily in the gaming [6] and fitness [11] industries. Often techniques that perform well inside the lab and within the boundaries of academic research do not provide the requisite goodness-of-fit for practical applications. See, for example, research [3] on the practical validation of fall detection algorithms. This may be due to the common problems faced when transferring any knowledge from the lab environment to the real-world. Real-world data is typically noisier and parts of it can be lost. Moreover, good results from a scientific paper may not always provide a practical solution when it comes to the real word. For example, assume the specificity of an activity alert algorithm is 99.9%; this means one false alert exists for every 1,000 normal observations. If we consider sliding windows of one second, with no overlapping, this same specificity translates into one false alert for every 17 minutes of normal activity. Clearly, this level of false alert is unacceptable for most applications.

To further focus the scope of our research, we address the detection of activities that involve some repetitive body or limb movements, and that last from several seconds to minutes and hours. We do not touch upon activities such as gesture recognition or fall detection, where a brief movement pattern occurs once and must be detected based on a single-pattern observation. This class of repetitive motion that we study covers many physical training activities, such as running, jogging, biking, and gym exercises. Numerous everyday life patterns are also periodic, such as tooth brushing and eating movement

patterns [1], or going up and down stairs. Additional applications can be found in industry and service environments, where employees often engage in undesired or sometimes prohibited activities of the described class. For example, in the workplace environment, running could be dangerous in certain areas, requiring that it be recognized and that safety regulations are enforced. Another example may be nutrition alerts that could be provided for people with dietary restrictions or those supposed to be fasting at certain times.

Repetitive movement patterns can also arise in the context of non-wearable applications that employ inertial sensors, as in the case of predictive maintenance. For example, detecting abnormal vibrations in industrial equipment could be a potential indication of some malfunction.

A mainstream approach to activity recognition in the examples above could be described as Activity Recognition Chain (ARC, see [5]). ARC involves the following data-processing stages:

- Preprocessing
- Segmentation
- Feature extraction
- Activity classification

The preprocessing typically involves unit-conversion, time-synchronization, and basic anomaly detection (for example, of sensor malfunctioning). This stage is of minor interest in this work since it usually does not involve deep research challenges. We address the three other stages in detail.

At the next stage, preprocessed data is partitioned into segments. The *sliding window* approach is recognized as the most straightforward and popular method for data segmentation. In this approach, a fixed time window of constant size is moved over the input time-series data. This allows a finite sample subset of the continuous input data to be extracted for further analysis. The window size typically varies from one to several seconds. This approach is especially relevant for the type of periodic activities we address in this work. For efficiency, this simple approach is sometimes complemented with additional algorithms to identify the beginnings of the "potentially interesting" interval in which the actual activity may take place. The same complementary techniques are also employed for non-periodic activity recognition, for example gesture recognition [9]. Although sliding windows technique is a mainstream one, research on optimal window size is rare. In addition, applications of sliding windows are typically based on pre-processed raw data from a <u>single</u> window, as described below.

Raw data that is generated by inertial sensors typically consists of a sequence of individual multi-attribute measurements, usually produced at a high frequency (e.g., 50 Hz or even 100 Hz). Hence, even a small window can give rise to a significant amount of data points that cannot be fed into a final activity classifier without additional processing and reduction of dimensionality. Consequently, in the next data processing stage, features are extracted from the data sample in each window considered. This transforms the data from a sequence of multi-dimensional measurements into a single feature vector. The extracted features can include both simple statistics (signal mean, minimum, maximum, standard deviation, mean crossing rate, etc.), and more complicated ones, such as smoothed output of frequency analysis based on Fast Fourier Transform (FFT), or wavelets. These features provide input to a final activity classification algorithm, to determine the activity class that corresponds to each window.

There seems to be no single dominating technique among the final activity classifiers. Numerous machine learning techniques, including decision trees, support vector machines (SVM), K-NN, random forests, neural networks, and Bayesian networks, are used at this final stage [5,10]. Their output constitutes the assignment of the considered window, as represented by its feature vector, to one of the activity or pattern classes. The class set should also include the so-called "null-class" (no important activity has been registered), which constitutes the majority of observations in many practical problems.

The framework described above raises numerous questions, such as: How do we decide which classifiers and learning algorithms are most appropriate for a specific problem? Which sliding window size is preferable? Which goodness-of-fit measures should be used to compare between different methodologies? The research agenda has been further exacerbated by the lack of public domain benchmark data to compare different approaches. Historically, video or speech recognition benchmark data sets are much more common. However, recently situation improved. Several sets of high-quality data from inertial sensors has become publicly available (see survey in [4]) and used as a basis for benchmarking.

In this paper, we focus on the problem of optimal window size, complementing it with a new dimension. We examine whether it is preferable to consider the usual method of individual time windows, or, alternatively, to extract data from several, possibly smaller consecutive windows, and then "vote between windows" for activity recognition.

RESEARCH CONTRIBUTION

We extend the conventional sliding-window approach to human activity recognition by introducing a family of enhanced recognition rules. We use several consecutive basic sliding windows to identify a certain activity class. Activity classifier is applied to each basic window. Ultimate activity detection is based on detection over a threshold number of basic windows (e.g., three-out-of-five). In some sense, this approach is similar to that of *ensemble learning* [12], which is currently gaining popularity in data mining. However, we perform voting-based decision-

making over several sliding windows, and not among several algorithms.

We evaluated these enhanced rules using one of the richest publicly available activity data set [4]. Our rules performed better than the conventional approach on a pool of eight special cases. These special cases arise from a combination of two popular classification algorithms, two testing approaches, and two different data sets. Given there are no generally accepted feature extraction and classification methods, this initial evaluation provides hope that our approach outperforms the conventional one for a broad range of activity classifiers and feature extraction algorithms.

Our enhanced rules, together with clearly specified parameters, potentially yield a framework that can determine optimal activity recognition rules given certain criteria for optimality and performance.

OPTIMAL WINDOW SIZE: STATE OF THE ART

As mentioned above, research on optimal window size is very scarce. Recent research by Banos et al. [2] provides an extensive review of the sliding window approach in human activity recognition. They conclude that there is a lack of methodological coherence with regard to window size determination. Their work also offers a good introduction to the aforementioned ARC stages.

Using a benchmark data set [4], Banos et al. [2] explore different window sizes ranging from 0.25 sec to 7 sec, using an F1-score[1] as the main goodness-of-fit measure. Contrary to prevailing assumptions in the field, they show that smaller window sizes do not necessarily imply poorer performance. Therefore, they recommend the use of 1 to 2 second windows. In other recent research, Fida et al. [7] recommend 1.5 second windows using another data set. Both papers conform to the abovementioned mainstream single-window approach. Our work challenges the optimality of this approach.

THEORETICAL APPROACH TO ENHANCED SLIDING WINDOW ACTIVITY DETECTION

We suggest the following four parameters for the enhanced sliding window approach:

- Basic sliding window length W
- Window overlap V
- Number of grouped basic sliding windows N
- Threshold number of adjacent basic windows M, $1 \leq M \leq N$, used to determine activity in a larger window that consists of a group of N basic windows

[1] F1-score is a single measure of accuracy capturing both precision and recall.

General feature extraction and classification techniques from the ARC chain can be incorporated into this approach. Let $\overline{D_n} = \{D_{n1}, \dots, D_{nL}\}$, $1 \leq n \leq N$, denote the sensor data set of an individual sliding window n out of N adjacent sliding windows. Parameter V determines the length of overlap between any two adjacent basic windows $\overline{D_n}, \overline{D_{n+1}}, 1 \leq n < N$, where $V=0$ signifies no overlap. Note that the data dimensionality L can be very large even for small-size windows. For example, 9 three-axis accelerometers with 50 Hz frequency provide 9*50*3=1,350 observations per second. Integrating a three-axis gyroscope would double this number.

Assume that some general feature extraction function $\bar{F}(\overline{D_n}) = (F_1(\overline{D_n}), \dots F_k(\overline{D_n}))$ extracts k features from the sensor data of the sliding window.

Feature selection procedures provide input to the activity detection classifier $A_n^C = A^C\left(\bar{F}(\overline{D_n})\right), 1 \leq n \leq N$. In this paper, we assume that function A is binary for any input assignment. That is, it either detects an activity of a certain class C during interval n: $(A_n^C = 1)$, or not: $A_n^C = 0$. Specific function A^C is typically related to some machine learning algorithm, such as decision tree, random forest, SVM, etc. Our methodology can be extended to detect multiple activities, as we discuss at the end of the paper.

The enhanced activity detection rule over a larger window, which is a union of N individual consecutive sliding windows, is defined as follows:

- Detect activity C over a series of N basic windows if it is detected in at least M intervals: $\sum_{n=1}^{N} A_n^C \geq M$
- Otherwise do not detect activity C

Mainstream sliding window approach is a special case of an enhanced one if $V = 0$, $N = 1$ and $M = 1$.

EVALUATION METHODOLOGY

Data Description

We used the public domain data set [4] for the initial evaluation of our approach. This data set has 33 activity classes, including walking, running, jumping, general fitness exercises, and specific body part activities. The data set also includes "zero-class" periods between exercises when no specific activity is performed. All activities were performed by 17 participants in the experiment. Nine inertial sensors were attached to each subject: 2 on the lower arms, 2 on the upper arms, 2 on the calves, 2 on the thighs, and 1 sensor on the back. Each sensor provided accelerometer, gyroscope, and magnetometer measurements. In this research, we address only accelerometer measurements.

In addition, the dataset incorporates three alternatives for sensor placement:

- Ideal placement by instructor

- Self-placement by user, which may accidentally deviate from an "ideal position"
- Intentional sensor displacement

In this paper, we address ideal placement and self-placement. The term "full data set" refers to the combination of the ideal and self-placement sets.

Using the enhanced activity detection rule, we attempted to recognize the following five activities, using the acceleration measurements from nine sensors.

- Running (activity 3 from [4])
- Running and jogging together (activities 2 and 3)
- Five different kinds of jumps (activities 4-8), or "aggregated jumps"
- Jumps with legs and arms opened and closed (activity 7)
- Frontal hand claps (activity 21)

In each of the five cases, we employed binary classification, to distinguish between a relatively rare activity (one of the above five activities) and all other activities, including no-activity intervals. The zero-class of our classification problem includes both fitness activities and periods between exercises.

Feature Extraction

For all nine sensors in each window, we extracted the basic features of acceleration: standard deviation, maximum and minimum values. In addition, we extracted FFT-based periodograms from sliding windows and computed smoothed periodogram values for nine frequencies: 1Hz, 1.5Hz, 2Hz, 2.5Hz,…, 5Hz.

Classification Methods

We applied two classification algorithms to the derived features: C5.0 decision trees and Breiman & Cutler's random forests, an ensemble learning technique based on decision trees. In both cases, we used implementations in the R programming language. In our experience, C5.0 decision trees outperform other decision tree algorithms, such as CART and CHAID, for this type of activity detection problems.

Training and Testing Sets

We considered two approaches to testing:

- Both training and testing data were randomly chosen from the same customer pool. Ten-fold random partitioning cross-validation was applied. We divided the data randomly into ten equal sets. Each set in turn was used for testing and the other 90% of data was set aside for training. The average goodness-of-fit for all testing sets was computed.
- Training was performed on 16 experiment participants. The last participant was used for testing. This procedure was repeated 17 times and average goodness-of-fit was computed.

Overall, we tested eight experimental configurations, manifested by the manipulation of two classification algorithms, two testing approaches, and two placement approaches (i.e., ideal placement vs. full data set).

Tested Parameters

We considered three sliding window lengths; W=5, 2.5, and 1 second. This initial evaluation did not include testing for window overlap (i.e., V=0). We tested the mainstream approach with M=1 and N=1 for each value of W. Our enhanced approach was tested for W=2.5 and 1 sec, with N=2 and N=5, respectively. M=1 and M=2 were tested for W=2.5 seconds. All values of M between 1 and 5 were tested for W=1 second. Table 1 summarizes all test cases.

Test-case	*W*	*M*	*N*
1	1	1	1
2	2.5	1	1
3	5	1	1
4	2.5	1	2
5	2.5	2	2
6	1	1	5
7	1	2	5
8	1	3	5
9	1	4	5
10	1	5	5

Table 1: Test cases employed in the evaluation

These parameter combinations enabled us to compare between the mainstream approach with a 5-second sliding window and several cases where the 5-second intervals were unions of smaller intervals.

Goodness-of-Fit-Metrics

For each test case, we computed several standard goodness-of-fit metrics for binary classification. The overall accuracy reflects the percentage of correctly identified observations. The specificity reflects the percentage of true negatives among all negatives (correct zero-class identifications among all zero-class observations). The sensitivity reflects the percentage of true positives among all positives (correct class C identifications among all class C observations). The precision reflects the percentage of true positives among all class C identifications (true positives and false negatives). Due to a very large null-class, the accuracy and specificity were typically close to 100%. Therefore, the precision and sensitivity were more informative than the accuracy and specificity.

In addition to the above metrics, we also used an F1-score, which balances between precision and sensitivity being equal to the harmonic mean of both. The F1-score was also used as the key single metric for comparing results, as depicted in Table 3 below.

EVALUATION RESULTS

Models with W=1 sec, N=5, and M=2 or 3 (test cases 7 and 8 from Table 1) demonstrated the best results among the enhanced approaches. Figure 1 compares their mean F1-averages for five activities with outputs of the three mainstream models: W=5, 2.5, and 1 second, respectively (test cases 1-3 from Table 1). Goodness-of-fit for other test cases was worse and we do not present it here.

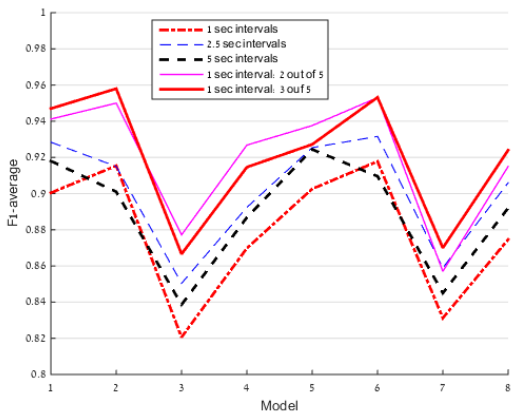


Figure 1. Graphical summary of results

Table 2 explains the settings of eight models laid-out on the x-axis in Figure 1.

Model	Classification	Testing on users	Data set
1	C5.0	Same	Full
2	C5.0	Same	Ideal
3	C5.0	Different	Full
4	C5.0	Different	Ideal
5	Random forest	Same	Full
6	Random forest	Same	Ideal
7	Random forest	Different	Full
8	Random forest	Different	Ideal

Table 2: Model description

Table 3 presents more detailed results of model testing with F1-scores per activity for main models in consideration.

DISCUSSION

First, consider the mainstream single-window approach as illustrated by all dashed curves in Figure 1. The 2.5-second windows achieved the best results. The 5-second intervals give slightly poorer results, and the 1-second intervals depicted, on average, the lowest F1-scores. These results are consistent with [2]: largest window size does not necessarily imply best goodness-of-fit.

However, both enhancement variations (two-out-of-five and three-out-of-five) demonstrated a uniformly better goodness-of-fit with respect to all conventional single-window approaches. There is no clear winner between two-out-of-five and three-out-of-five enhancements.

There is also no clear winner between C5.0 and random forests. Both methods demonstrate similar performance, with C5.0 slightly better on average. It is important to note that the enhanced approach prevails for two different advanced classification methods.

As expected, an ideal placement data set, on average, implies better performance than a full data set. However, the differences are small when testing is performed on the same users. In contrast, testing on different users implies significantly better results for the ideal sensor placement. We also observed that the average goodness-of-fit for testing on different users is derived from the good fit for most of them and is an unsatisfactory fit for a smaller set (typically, 2-4 users out of 17).

Reviewing detailed results in Table 3, we observe that aggregated jumps were detected with higher accuracy than a specific jump type. In fact, the aggregated jump type was detected better than any other activity. This result is intuitive; overall, jumps constitute a very distinctive type of activity; however, difference between specific types of jumps is significantly smaller.

CONCLUSIONS AND FUTURE WORK

We found that using voting across consecutive windows is superior for recognizing activities than a conventional single-window method. Most prominently, this also holds true when comparing between a single interval and its breakdown into a set of smaller fragment intervals (e.g., see 5 second single window versus three-out-of-five). Therefore, using our new method does not necessarily come at the cost of longer response times. Our findings are of course amenable to further sensitivity testing and optimization as discussed next.

Optimization Framework

The approach described above uses four parameters: basic sliding window length, window overlap, number of grouped windows, and detection of the threshold number of windows in the group. We presented very promising initial results for this approach. However, we do not describe how to choose the best parameter combination for specific practical applications. This requires first determining the "goodness-of-fit" criteria.

For example, a reasonable criterion may use False Positive Rate (FPR, or specificity) per chosen time duration and some acceptable threshold, while aiming to maximize the portion of True Positive Rate (TPR, or sensitivity).

This may be formulated as follows:

$$\begin{cases} \max\left(E\left(TPR(V,W,M,N)\right)\right) \\ s.t.: E(FPR(V,W,M,N,T)) \leq \alpha \end{cases},$$

where T and α define time-based constraint on FPR. For example, T=40 hours and α=1 designate the constraint: "no more than 1 false alert on average during 40 hours".

Finding efficient ways to run and validate this framework is an important research challenge. It could also give rise to

more "objective" ways of activity detection design and fine tuning.

Generalization to Several Features

In this paper, we considered a binary classification problem, which is appropriate, for example, to detect a single activity. In contrast, many practical settings demand the detection of a number of different activities. For example, consider an "exercise profile" in a gym that shows the durations of different exercises and determines whether an exercise was performed correctly. Our approach can be generalized to a multi-activity case by using multi-class machine learning classifiers (for example, C5.0 and random forests that were used above). However, if an enhanced alert rule is activated in the case of "minority voting" (two-out-of-five intervals, for example), reasonable tiebreaking strategies have to be designed for the case when two different activities are detected during a large window.

Other Feature Extraction and Classification Methods

The enhanced sliding window approach should be validated on a larger set of feature extraction and final classification methods. SVM and K-NN are the first obvious candidates to be considered for further evaluation. It is also worth examining whether the use of gyroscope measurements could potentially improve goodness-of-fit.

ACKNOWLEDGMENTS

We thank the authors of [4] for the publicly available benchmark data set that we used to evaluate our methods.

REFERENCES

1. Amft, O., Junker, H. and Tröster, G. 2005, October. Detection of eating and drinking arm gestures using inertial body-worn sensors. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium* (pp. 160-163). IEEE.
2. Banos, O., Galvez, J.M., Damas, M., Pomares, H. and Rojas, I. 2014. Window size impact in human activity recognition. *Sensors*, *14*(4), pp.6474-6499.
3. Bagalà, F., Becker, C., Cappello, A., Chiari, L., Aminian, K., Hausdorff, J.M., Zijlstra, W. and Klenk, J. 2012. Evaluation of accelerometer-based fall detection algorithms on real-world falls. *PloS one*, *7*(5), p.e37062.
4. Banos, O., Damas, M., Pomares, H., Rojas, I., Toth, M.A. and Amft, O. 2012, September. A benchmark dataset to evaluate sensor displacement in activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 1026-1035). ACM.
5. Bulling, A., Blanke, U. and Schiele, B. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, *46*(3), p.1-33.
6. Darby, J., Li, B. and Costen, N. 2008. Activity classification for interactive game interfaces. *International Journal of Computer Games Technology*, *2008*, p.11.
7. Fida, B., Bernabucci, I., Bibbo, D., Conforto, S. and Schmid, M. 2015. Varying behavior of different window sizes on the classification of static and dynamic physical activities from a single accelerometer. *Medical engineering & physics*.
8. Igual, R., Medrano, C., and Plaza, I. 2013. Challenges, issues and trends in fall detection systems. *Biomed. Eng. Online*, *12*(66), pp.1-66.
9. Junker, H., Amft, O., Lukowicz, P. and Tröster, G. 2008. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recognition*, *41*(6), pp.2010-2024.
10. Qian, H., Mao, Y., Xiang, W. and Wang, Z. 2010. Recognition of human activities using SVM multi-class classifier. *Pattern Recognition Letters*, *31*(2), pp.100-111.
11. Velloso, E., Bulling, A., Gellersen, H., Ugulino, W. and Fuks H. 2013, March. Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference* (pp. 116-123). ACM.
12. Zhang, C. and Ma, Y. 2012. *Ensemble Machine Learning*. Springer.

Classifier	Type of testing set	Segmentation	F1 scores					
			Running	Running and jogging	Aggregated jumps	Specific jump	Hand claps	Average
C5.0 decision tree	Same users: full data set	1 sec	92.0%	92.8%	96.1%	85.0%	84.3%	90.0%
		2.5 sec	93.9%	94.3%	96.7%	89.3%	90.0%	92.8%
		5 sec	93.7%	93.7%	96.5%	89.8%	85.3%	91.8%
		1 sec: 2 out of 5	95.4%	94.7%	97.1%	91.9%	91.5%	94.1%
		1 sec: 3 out of 5	96.9%	96.3%	98.2%	90.9%	91.2%	94.7%
	Same users: ideal data set	1 sec	92.4%	93.5%	97.5%	88.0%	86.3%	91.5%
		2.5 sec	93.0%	93.9%	99.1%	87.2%	84.3%	91.5%
		5 sec	92.7%	93.8%	98.4%	86.2%	79.2%	90.1%
		1 sec: 2 out of 5	96.4%	94.9%	98.2%	91.5%	94.1%	95.0%
		1 sec: 3 out of 5	97.5%	95.8%	98.7%	95.0%	92.2%	95.8%
	Different users: full data set	1 sec	76.2%	86.6%	94.8%	80.1%	72.6%	82.1%
		2.5 sec	83.9%	87.8%	95.7%	80.0%	77.8%	85.0%
		5 sec	74.9%	83.6%	96.4%	82.7%	81.7%	83.9%
		1 sec: 2 out of 5	84.1%	90.8%	96.2%	85.1%	82.4%	87.7%
		1 sec: 3 out of 5	82.7%	91.2%	96.9%	86.2%	76.3%	86.7%
	Different users: ideal data set	1 sec	83.4%	89.7%	97.4%	84.3%	80.1%	87.0%
		2.5 sec	87.2%	90.1%	98.2%	86.3%	84.4%	89.2%
		5 sec	84.0%	87.8%	99.1%	89.8%	82.7%	88.7%
		1 sec: 2 out of 5	88.3%	93.4%	98.0%	89.8%	93.9%	92.7%
		1 sec: 3 out of 5	90.4%	92.6%	98.9%	91.1%	84.3%	91.5%
Random forest	Same users: full data set	1 sec	93.5%	94.1%	96.1%	87.8%	79.8%	90.3%
		2.5 sec	97.1%	95.1%	97.6%	91.8%	81.1%	92.5%
		5 sec	95.6%	95.3%	97.4%	93.9%	80.0%	92.4%
		1 sec: 2 out of 5	95.8%	95.4%	96.6%	91.1%	89.9%	93.8%
		1 sec: 3 out of 5	95.5%	96.1%	97.2%	91.0%	83.8%	92.7%
	Same users: ideal data set	1 sec	94.6%	95.0%	97.9%	90.5%	80.9%	91.8%
		2.5 sec	97.5%	96.1%	98.5%	94.5%	79.2%	93.2%
		5 sec	96.0%	95.3%	98.7%	95.9%	68.9%	91.0%
		1 sec: 2 out of 5	96.6%	96.1%	98.4%	93.4%	92.0%	95.3%
		1 sec: 3 out of 5	97.2%	96.7%	99.1%	93.9%	89.7%	95.3%
	Different users: full data set	1 sec	80.7%	90.1%	95.5%	81.9%	67.4%	83.1%
		2.5 sec	85.0%	90.2%	96.8%	84.9%	72.4%	85.9%
		5 sec	79.8%	89.2%	96.9%	88.3%	68.4%	84.5%
		1 sec: 2 out of 5	80.3%	88.7%	95.1%	84.4%	80.2%	85.7%
		1 sec: 3 out of 5	84.3%	92.0%	96.7%	86.2%	75.8%	87.0%
	Different users: ideal data set	1 sec	86.1%	90.7%	98.0%	88.2%	74.3%	87.5%
		2.5 sec	91.6%	90.5%	98.5%	95.8%	76.5%	90.6%
		5 sec	85.9%	90.9%	98.0%	94.1%	77.1%	89.2%
		1 sec: 2 out of 5	87.1%	90.8%	98.5%	91.3%	89.8%	91.5%
		1 sec: 3 out of 5	89.9%	92.3%	98.9%	92.4%	88.3%	92.4%

Table 3: Summary of F1-scores for different models and data sets

SMILEY: Emotion Therapy through a Smart-Scarf

Chen Guo, Yingjie Victor Chen, Zhenyu Cheryl Qian, Yue Ma, Hanhdung Dinh, Saikiran Anasingaraju
Purdue University, West Lafayette, IN 47906
{guo171, victorchen, qianz, ma173, hdinh, sanasing}@purdue.edu

ABSTRACT
Group interactions are very common for human beings. When a group is in discussion or collaboration, not only the physical environment, but also the surrounding emotional environment will affect group interaction process. To create an emotional healthy atmosphere among group members with an ambient approach, we designed a wearable product called *Smiley*. It is a smart-scarf that can adjust its color automatically according to the emotion statuses of both wearer and viewers. Research on color has shown that the environment can prominently impact a human's emotion. We applied this theory to a scarf design, using a dynamic surface to adjust people's emotional state. The evaluation shows that *Smiley* is a physically and socially acceptable wearable product and has potentials to improve human emotional health.

Author Keywords
Smart object; aesthetic interaction; emotion regulation; color combination; wearable interaction.

ACM Classification Keywords
H.5.m. User Interfaces: Miscellaneous.

INTRODUCTION
Wireless sensor technology has now made it possible to track, manage, and analyze body data, and intelligent products are integrated into our daily lives with dynamic interactions. It is crucial to design smart objects for improving human emotional health.

While in a group, an individual may find that emotion and behavior can be significantly influenced, motivated, or even controlled by the social environment that the group members provide. Brides and grooms often break into tears at weddings. Fans can lose control at a sports event. Today's technologies have made it possible to track and analyze body data to detect a person's emotions and moods. A potential exists to employ these data to achieve a human's self-consciousness and emotional awareness, positively influencing his/her moods and improving psychological health.

To improve emotional health among group members, we propose a tangible interaction wearable interface, the *Smiley*. It's a smart-scarf that can change color automatically according to the emotions of wearers and viewers. The scarf has a dynamic surface that helps its wearer adjust his/her emotional state and enhances the social bonds between people. If the user is alone, *Smiley* will change its color to cheer up the user or calm him/her down. If a group of people wearing *Smileys* are at a party together, the *Smileys* will change their colors to cheer up those in sight as based on the severity-of-emotion index. Emotional data are continuously transmitted through the human body from inside to outside, from private to public. It also becomes the necessary medium of our being and self-presentation in the world.

RELATED WORK

Aesthetic Interaction and Body Experience
Wearable smart products, such as watches, glasses, and jewelry, have made technology pervasive by interweaving it into daily life. We are interested in not only getting the data from the human body, but also influencing it with a positive and ambient approach. A tendrils garment was a responsive kinetic wearable artwork that reacted to shared active touch both locally on the garment and collectively through a remote networked wearable armband [1]. The design concept and aesthetic interaction of this tendrils garment is the source of inspiration to our design principle. The purpose of our project is to allow wearers to be aware of negative emotions in their daily lives and also in those of others by offering a combination of colors as a solution.

Emotion and Color Combinations

Researchers in Korea designed a color combination emotion model [2]. It consisted of a three-dimension axis that represents a soft-hard parameter, a light-heavy parameter, and a splendid-sober parameter. Nine emotion groups are matched with the color combinations. In Japan, Shigenobu Kobayashi analyzed a variety of color combinations with references to different emotion adjectives in his book *A Book of Colors* [3]. The main concentration of these researches is interior design, and it inspired us to explore the relationship between color combination and emotion.

Emotion and Technology

IUI 2016 Workshop: Interacting with Smart Objects, March 10th, 2016, Sonoma, CA, USA

Some researchers provide a theoretical and empirical rationale for the use of heart rate variability (HRV) analysis to measure regulated emotional responding [4]. HRV measures heart rate changes over time. Positive stimuli are correlated with an increase in HRV, and negative stimuli are correlated to a decrease in HRV. Thus we can use heart rate variability to measure people's emotional states. Researchers in the mixed-reality lab (MXR) at City University London explored how to achieve a rapid color-changing wearable display [5]. They created a non-emissive analog fabric display called AmbiKraf, and the technology was based on embedded semiconductor peltier junctions and thermochromic inks. The controller can control the temperature in the soft fabric and enable the thermochromic ink to display different colors [5]. The AmbiKraf project inspired us to find an embedded non-emissive way to change wearable products.

USER-CENTERED DESIGN PROCESS

Our concept arose from a series of user studies while we were seeking solutions to use body data to influence a user's mood. With a snowball-sampling method, we recruited 12 participants to describe their emotion experience. Participants were all college students and staff members. Their ages ranged from the 20s to the 40s. They were interviewed to describe what their bodies had experienced in terms of depressed situations or overexcited situations and what contexts have typically affected their experience of the phenomenon in a group environment.

To set up the guidelines of our design, we found that people want to change their emotional state in a subtle way instead of revealing their emotions in public. Participants also revealed that group environment greatly affected their emotions, especially for overexcited emotion.

We brainstormed and sketched several ideas of the concept. All initial designs were presented to the same participants recruited in the user study. The idea of a smart hair band, a bracelet, a vest, and a scarf received positive feedbacks from participants. However, they preferred to wear a scarf because it wouldn't draw too much attention and could be worn every day.

OUR SOLUTION: AN EMPATHIC TOOL – SMILEY

Based on users' feedback, we proposed a tangible interaction interface and designed a smart-scarf prototype that we call *Smiley*. Our concept is to change the scarf's color to help the wearer control his or her emotion and enhance the social bonds among people in a group environment in an aesthetical and subtle way.

Interaction Concept

Our research aims to help people become more and more aware of their mental well-being and improve their emotional health. We try to explore the connection between body and mind, interpersonal and intrapersonal communication, and the interaction concept is developed into three scenarios: self, one-to-one, and one-to-many.

Self: The first scenario is based on the concept of self-therapy (Fig. 1). When the user is alone, *Smiley* will detect his/her emotion to moderate self-emotion. When the user is wearing the scarf, it may be hard to notice the color. Thus *Smiley* can emit different scents to make people feel good as it helps the user by scent as the color changes. If the user is sad, *Smiley* will emit the scent of jasmine and change its color to cheer him/her up. Similarly, when the user is overexcited, *Smiley* will emit the scent of lavender and change its color to calm him/her down. When the user is in a calm and positive mood, *Smiley* will maintain a delightful color to match the wearer's cloth and emit the scent of jasmine. The color choices are determined by our color combination research.

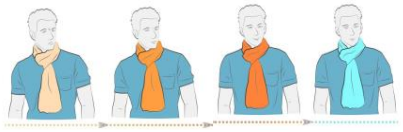


Figure 1. Scenario of self-therapy

One-to-one: This scenario is based on the concept of social emotion therapy. When two users interact, *Smiley* on one side will become the therapy for the other side because little cognitive effort is necessary to notice the scarf color on the other side (Fig. 2). *Smiley* will receive the emotion data from the other side and adjust its color with the hope of bringing warm, positive, and motivated feelings to each other. The color choices of emotion therapy are determined by our color combination research. Thus it is possible to cheer up a sad person on one side and calm down an overexcited person on the other side at the same time. During the whole conversation, both scarves will continue adjusting color to make both users feel warmly good, but not overexcited.

Figure 2. Scenario of interaction therapy

Group: It is possible to extend *Smiley*'s application to a group setting. When several *Smiley* scarves are in a group, such as at a party, the collected emotion data can be shared, exchanged, and integrated. On one hand, a *Smiley* wearer can be more sensitive to the emotion conditions of others, initiating and changing topics to warm up or calm down the condition. On the other hand, *Smiley* scarves can be collectively grouped into several clusters and targeted to motivate several people simultaneously.

In a group, *Smiley* will collect an emotional index from any user who can see it within its range. *Smiley* will prioritize the emotional index based on the severity of the group's current emotional state based on the unified data collection, with negative emotions having top priority. After targeting

the most negative emotion user, *Smiley* will change its color to create a new environment for the second user. *Smiley* will continue monitoring to show a mild warm look, creating a happy environment, but not one that is too wild. Most of the time a person may see only the people on the other side, being unable to see those beside him/her. Thus *Smileys* may be automatically grouped into several clusters to motivate several people simultaneously.

Interface Design

Our user research finding showed that the wearable product needs to be flexible, lightweight, comfortable, conductive, and reasonably priced. We made our design to attach a heart rate sensor module, a skin conductance sensor module, a color sensor module, and an IR transceiver to a LilyPad Arduino (Fig. 3). The electronic sensor developed by the University of Sussex Innovation Centre can remotely detect heart rate from one meter away without being physically attached to the human body [5]. Researchers reported that heart rate variability and skin conductance response were the two most reliable physiological signals to monitor emotions, including sad, calm, happy, angry, and stressful [6][7]. Thus using heart rate sensors and skin conductance sensors are enough to detect depressed, overexcited, and calm emotions. The scent capsules inside the scarf will emit different scents to help wearers be aware of the color changes when they are alone.

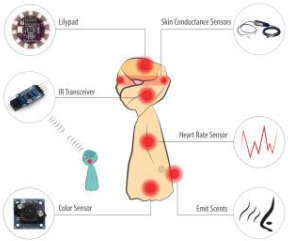


Figure 3. Interface Diagram of a *Smiley*

Data transfer is the key for the interpersonal interaction part of the interface, which enhances social bonds among its users. An infrared receiver is capable of sending and receiving data. Since people must see the scarf to be affected, we chose IR because it is directional. As a receiver, IR is also able to receive other heart rate data from viewers' scarves for sending to the controller.

The LilyPad controller receives and computes the raw data sent by sensors and sends the signal to control the display. LilyPad receives heart beat rate, computes the current emotion index, and then scans the surroundings to see if any other users are in sight. If not, the scarf will work in self-mode. If it detects other *Smiley* scarves in sight, it will then work in the one-to-one or one-to-many mode.

The dynamic wearable scarf can change its color and influence emotional states to embody emotions. We will choose cotton, thermochromic ink, conductive yarn, and conductive fabric to make the scarf. Dyed with thermochromic inks, the conductive fabric contains interwoven stainless steel yarns as the heat source. Electric activity heats up the conductive yarns, and the heat then activates thermochromic inks. These inks have the capability of changing a color to colorless by heating. Thus we can create a colorful scarf using CMY (cyan, magenta, yellow) colors (Fig. 4). By turning each color block on or off, we are able to show different colors with different variations. The temperature is controlled by electrical conductors, thereby controlling the color of the scarf. Thermochromic ink will vanish when the temperature is around 31ºC, which is slightly higher than indoor ambient temperature in the winter. But it is durable and acceptable for the human body.

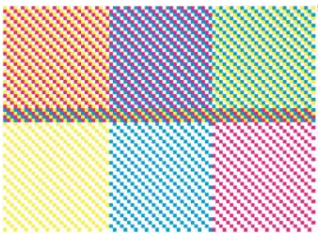


Figure 4. Smiley color variation examples: turn on/off each color block of CMY to show different colors

We intend to match the color of the wearer's scarf with the blouse or shirt that he or she is wearing so that the color combination will have a positive effect on human emotions in a subtle way. The vendor Keyence provides a small-enough RGB optic sensor that can be embedded into our scarf [8] to measure the color of the cloth. We collected the color combination details from *A book of colors* [3] and formulated the collected data into a spreadsheet that shows a one-to-one relation between a color combination and its effect. The first part of the spreadsheet gives the color combinations and their effects in general, and the second part explains the emotional effects of the combinations color-wise. Even though we formulated a big list of color combinations, we could use only the combinations for seven basic colors because of time constraints of the first iteration of this product design (Fig. 5). Combing the severity of emotional states with the color sensor data, the LilyPad will determine the display color and send corresponding electric currents to the scarf.

Figure 5. *Smiley* color combination examples

USER EVALUATION

The prototype is not actually functional. As a formative evaluation of *Smiley*, we conducted a Wizard of Oz study with the think-aloud method. This kind of study enables us to test the concept and techniques of the prototype [9]. A researcher, called the wizard, simulated the response of the scarf interface and tested the design concept.

Study Setup

We recruited 11 participants to conduct usability testing. Participants were all college students and staff members ranging in age from the 20s through the 40s. Seven are males and four are females. They have different backgrounds and nationalities (U.S., India, China, Iran, and Vietnam).

We divided participants into four different groups with one, two, three, or four participants. Each group spent approximately 20 minutes to perform tasks. All the participants were assigned a scarf, and the scarf color matched the colors of their clothes. First, we carefully explained the prototype and the testing procedure. Second, we asked the participants' current feelings. A researcher then joined the group and simulated the color change of the scarf in real time based on the extreme emotions in this group. At the end of the study, we asked participants several structured questions and measured how well they responded to six areas: efficiency, affect, control, helpfulness, learnability, and emotional response. Product evaluation was based on the Likert scale from 1 to 5: strongly disagree, disagree, neutral, agree, and strongly agree. The open questions we asked included these: Can you describe your feelings before the scarf changes color? Can you identify your current mood? What affected your emotions and how were they influenced? What do you feel about the scarf's color changes?

Problems Found and Design Plans to Improve Them

Our evaluation indicated that the smart-scarf helped users control their emotions (average 4), and it was easy to use (average 4). They felt pleasure (average 4), comfortable (average 3), and engagement (average 4) wearing the scarf. They were satisfied with the aesthetics (average 4) of the scarf. However, we did identify problems of *Smiley* and proposed design suggestions to improve the interface.

Add gender and culture consideration in the design

Color perception is influenced by gender and culture. Depending on individual's culture, gender, background, personal experience, people have different feelings about different color combinations. For instance, a yellow shirt with a blue scarf reminded Indian participants of the Diwali festival; American participants didn't like the combination because they considered it girlish. We should customize the scarf based on gender and culture differences toward the use of colors.

Add control to the scarf

How do you control the scarf? Is it always desirable to indicate specific emotions in a group? Sometimes people want to hide their real emotions instead of expressing them. We should allow wearers to shut down the smart-scarf if they don't want to be exposed in public.

CONCLUSION

Smiley is designed to gain understanding of self- and emotion-consciousness, facilitate communication among people, and enhance social bonds in a group environment. To improve emotional health, we provide a concept to combine smart textiles and soft electronics to create a smart color-changing scarf. Although new technology is often rapidly adapted, it is still difficult for people to use smart garments in their daily life. Our usability testing results shows that fulfilling such user needs as privacy and intimacy need to be considered. We should also address the gender and culture differences in the next design stage.

REFERENCES

[1] Schiphorst, T. and Seo, J. Tendrils: exploring the poetics of collective touch in wearable art. In *Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction*, ACM Press (2011), 397-398.

[2] Lee, Y.J. and Lee, J. The development of an emotion model based on colour combinations. *International Journal of Consumer Studies 30*, 2 (2006), 122-136.

[3] Kobayashi, S. *A Book of Colors: Matching Colors, Combining Colors, Color Designing, Color Decorating*. Oxford University Press, 1987.

[4] Appelhans, B.M. and Luecken, L.J. Heart rate variability as an index of regulated emotional responding. *Review of general psychology 10*, 3 (2006), 229-240.

[5] Peiris, R.L., Tharakan, M.J., Fernando, N. and Chrok, A.D. AmbiKraf: a nonemissive fabric display for fast changing textile animation. In *2011 IFIP 9th International Conference on Embedded and Ubiquitous Computing (EUC)*, IEEE Press (2011), 221-228.

[6] Harl, C.J., Prance, R.J. and Prance, H. Remote monitoring of biodynamic activity using electric potential sensors. In *Journal of Physics: Conference Series 142*, 1 (2008), 012042.

[7] Healey, J.A. and Picard, R.W. Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on 6*, 2 (2005), 156-166.

[8] Keyence Corporation. More stability & reliability in every detection The Smartest RGB Sensor in the Industry. Keyence, 2004.

[9] Ferreira, P., Sanches, P., Höök, K. and Jaensson, T. License to chill!: How to empower users to cope with stress. In *Proceedings of the 5th Nordic Conference on Human-computer Interaction: Building Bridges*, ACM Press (2008), 123-132.

Learning User Intentions Spanning Multiple Domains

Ming Sun **Yun-Nung Chen** **Alexander I. Rudnicky**
School of Computer Science, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213, USA
{mings, yvchen, air}@cs.cmu.edu

ABSTRACT
People are able to interact with domain-specific applications in smart environments and get assistance with specific tasks. Current intelligent agents (IAs) tend to be limited to specific applications. In order to engage in more complex activities, users have to directly manage a task that may span multiple applications. An ideal personal IA would be able to learn, over time, about these tasks that span different resources. This paper addresses the problem of multi-domain task assistance in the context of spoken dialog systems. We propose approaches to discovering users' higher level intentions and using this information to assist users in their complex tasks. We collected real-life smart phone usage data from 14 participants and investigated how to extract high-level intents from users' descriptions of their activities. Our experiments show that understanding high-level tasks allows the agent to actively suggest apps relevant to pursuing particular user goals and reduce the cost of users' self-management.

Author Keywords
Multi-domain; User intention; Spoken dialog system (SDS); Intelligent assistant (IA); Language understanding.

ACM Classification Keywords
I.2.1 Applications and Expert Systems: Natural language interfaces; I.2.7 Natural Language Processing: Language Parsing and Understanding; I.2.11 Distributed Artificial Intelligence: Intelligent agents

INTRODUCTION
Environments, such as a home, can host smart objects/devices where each typically operates in a specific domain (for example, climate control or security). Each such object, by design manages few domains, usually one. For example, a fridge may support a grocery domain by tracking vegetables inside of it, perhaps additionally helping to compose a shopping list. It might even be configured to support sharing of information to other domains known to it, but such functionality would not be scalable and might lack potentially desirable adaptive features.

IUI 2016 Workshop: A joint Workshop on Smart Connected and Wearable Things, March 10th, 2016, Sonoma, CA, USA

urn:nbn:de:tuda-tuprints-54208

In contrast, users often mentally arrange tasks that span domains and manage the information shared among them. However, even if we assume that environment information is aggregated into a handheld device (e.g., phone) in the form of apps, the process of launching apps one by one may be time-consuming and difficult for users, especially for elders and ones with (visual) disabilities, although vocabularies of a touch-screen or gestures have been enriched significantly over the past decade [10]. We would want our personal intelligent agents (IAs) to automatically help users organize tasks across domains (or, apps) given a user's request expressed, in language, at the level of intentions. For example, upon receiving a request like "can you help me plan a grocery shopping trip?" the IA should determine what foods are out of stock (FRIDGE), the next bus to a nearby supermarket (NAVIGATION) and finally the locations of the food inside the supermarket (AISLEFINDER).

Conventional dialog systems operate in specific domains such as restaurant selection [31, 11], transit information [21] or event arrangement [19]. Multi-domain dialog systems have been studied in the past [16, 22, 15, 18, 14, 4, 24, 5, 7, 6], but they typically lack the capability of understanding the user's goal or high-level intention. As a result, such system has certain drawbacks (see example in Fig 1): 1) it passively selects *one* domain at a time given the user request; 2) it has no expectation of the next domain; 3) it does not maintain a shared context across domain boundaries. Consequently, the multi-domain conversation between the agent and the user will not be as natural/efficient as a human assistant would provide.

The IA should assist human users in interacting with multiple domains, as in the following use cases: 1) during conversation, our model can use the current context to predict the user's next action as well as his high-level intention [27]; 2) the IA should understand the user's high-level intention stated at the beginning of the conversation and coordinate existing domains to accomplish this intention [28, 29]. In this paper, we summarize our findings from our previous work, followed by the remaining challenges in this multi-domain agent setup.

DATA COLLECTION
To let the agent learn how human users coordinate existing functionalities for complex tasks, there are two ways: 1) users can explicitly instruct/author such procedural knowledge [23, 1]; 2) the agent can observe how users perform such tasks and learn. The first approach may rely on the agent's capability to comprehend sequential instructions, as well as user' capability to thoroughly define a task via language. In this work, we take the second approach. We use a smartphone to allow us to investigate cross-domain task management without having to deal with the complexity of a fully situated implementation.

S: What can I do for you?

U: Could you arrange a dinner for me and my friends?

S: Sorry I don't understand that. What can I do for you?

U: Can I book a table for three in Tākō downtown for this Friday?

...

S: OK. What can I do for you next?

U: Show me the bus from here.

S: Where is your destination please?

U: Tākō downtown Pittsburgh.

...

S: What should I do next?

U: Send the bus route to Carrie and Peter.

S: OK. Sending the following message to Carrie and Peter: "the bus route".

A: What can I do for you?

U: Could you arrange a dinner for me and my friends?

A : What kind of food do you prefer?

U: Mexican?

A : How about Tākō? I can book a table for you.

U: Sounds good! Can I take a bus there?

A : 61 A/B/C/D can take you there. Do you want to send this to your friends?

U: Great! Send it to Carrie and Peter.

A: OK. The bus route 61 has been sent.

Figure 1. Left: example dialog between user (U) and a classic multi-domain dialog system (S); Right: example dialog between user (U) and human assistant (A).

We logged real-life interactions at app-level (app invocation + when + where), segmenting a day's log into episodes (as separated by periods of 3 minute inactivity). Each episode could contain more than one app invocation. We asked users to annotate each episode to: 1) group apps used for a particular goal; and 2) describe the goal in language. Meta information such as day, time, location was shown to the user to aid recall. Users were asked to re-enact the smart phone interaction by speaking with a Wizard-of-Oz dialog system. The participants were not required to follow the order of the applications they used on the smart phones. Other than for remaining on-task, we did not constrain expression. The wizard (21-year-old male native English speaker) was instructed to respond directly to a participant's goal-directed requests and to not accept out-of-domain inputs. An example of annotation and Wizard-of-Oz dialog is shown in Figure 2.

We had 14 participants and collected 533 sessions; mean age for the 4 male participants was 23.0 and 34.6 for the 10 females. 12 were native English speakers. On average, each user interacts with 19 different apps. Across 14 users, a total of 132 apps were used. Details of the collection are provided in [25].

DOMAIN TRANSITION

As mentioned earlier, users can mentally coordinate a set of domains to accomplish complex tasks. However, this would require user to manually launch the next domain through speech, touch-screen or other modalities. Ideally, we would like the agent to have some expectation of the follow-up domain such that the transition can be smooth and easy for the user. In our previous work, we built context-based model to predict the next domain a user would interact with [27]. The agent could use this expectation to warm up the predicted app in the background, forward current information about the interaction so far to this app, or even proactively fetch the information from the next app and offer to the user.

Meta: 20150203; Tuesday; 10:48; Home

Apps: settings; music; mms

Desc: play music via bluetooth speaker

User: Connect my phone to bluetooth speaker.
Wizard: Connected to bluetooth speaker.
User: And play music.
Wizard: What music would you like?
User: Shuffle the playlist.
Wizard: I will play the music for you.

Figure 2. User annotation: 1) user connected SETTINGS and MUSIC; and 2) user noted that these two apps were used to *play music via bluetooth speaker*. Wizard-of-Oz dialog was collected and manually transcribed.

Table 1. Ranked features for prediction tasks without goal statement. Meta is a combination of time, day and location.

Rank	App Prediction	Intention Prediction
1	*Lang+App*	*All*
2	*All*	*Meta+App*
3	*Meta+App*	*Lang+App*
4	Lang	App
5	App	Lang
6	Majority	Location
7	Time	Majority
8	Day	Meta
9	Location	Time
10	Meta	Day

The contexts for prediction include 1) time (hours based on 24 hour clock), day (Monday, Tuesday, ..., Sunday), location (street names e.g, "Forbes Ave"); 2) previously launched app; 3) (noun and verb) words in user utterance. The intuition behind these contexts are: 1) people would have different tasks in different places or time; 2) what information people obtained (in certain app via certain speech command) would indicate what he might want to seek next. From our experiment, we found that previous app and user utterance are very informative by using multi-class (i.e., app ids) classification models. The rank of the result is shown in the left side of Table 1.

INTENTION PREDICTION

Users might not explicitly express their ultimate goals or intentions (e.g., "plan a dinner with friends"). The agent might need to infer the user's ultimate intention to provide timely assistance. Knowing that the user's intention is to "schedule a meeting", it could refer to the common procedure of this user or others to accomplish this task. Note that sharing the agent's state of understanding can improve the communication channel transparency: The agent can say "I think you want to *plan a dinner*. Let's find a restaurant for you first." In this way, the agent can reveal its (mis-)understanding of the task at hand, allowing the user to correct it when needed.

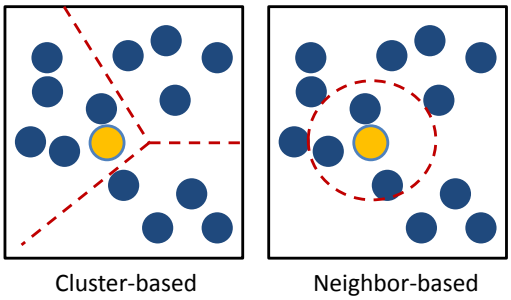


Figure 3. Cluster-based vs. Neighbor-based intention definition.

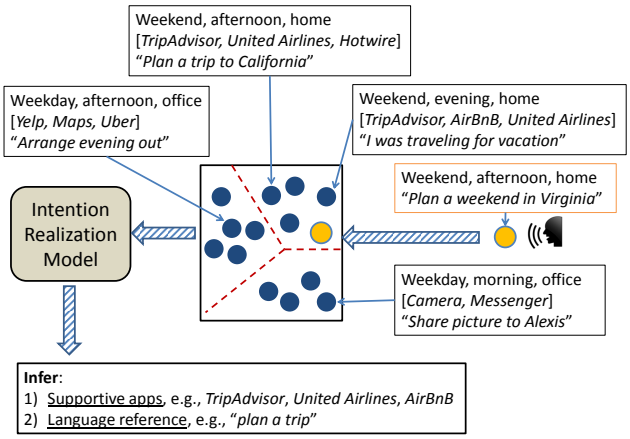


Figure 4. Example of cluster-based intention understanding and realization. Solid nodes denote past interactions (blue) and current input (yellow).

There are two cases in predicting a user's high-level intention. First, when the user does not explicitly state his intention, the agent can infer it by context (time, day, location, apps and user utterances). On the other hand, the user may directly express the intention via language at the beginning of the interaction. The agent could map that expression to a certain intention. Regardless of the different use cases, this process of inferring user's intention has two phases: 1) define a finite set of intentions and 2) recognize the intention given the input (either the context or the user's utterance at the outset).

Intention Definition

We let our agent automatically cluster seen interactions into K_C groups, each representing one intention. Bag-of-words features extracted from each interaction are words (lemmatized nouns and verbs) in user-generated language (i.e., task description and user utterances) as well as other contexts. The number of clusters K_C can be automatically optimized by using gap statistic [30]. We call this cluster-based intention definition.

We conducted a user study to examine the effectiveness of this cluster-based approach [28]. Users were shown the cluster members, each with the produced dialog, apps involved, task description provided by user. We asked each user to rate their agreement with the statement that the tasks shown in each cluster are essentially of the same nature. On average, we obtained 4.2 out of 5.0, where 5.0 indicates strong agreement. An example of clustered user tasks are shown in Table 2.

Similarly, we can use the K_N nearest neighbors of the input to denote the current intention (neighbor-based definition). The difference is illustrated in Fig 3. In this approach, K_N was set to the square root of the number of training examples [8]. The advantage of cluster-based approach is that the agent has awareness of the typical tasks the user performs everyday. This can be useful in the future when the agent lists a few possible intentions for grounding, to prevent potential misunderstandings.

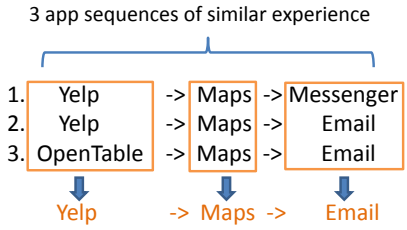


Figure 5. Example of Representative Sequence approach

Intention Recognition

In this section, we introduce two use cases of intention recognition: one with ultimate goal/intention expressed and one without. When the user does not express his goal, the agent needs to recognize the goal implicitly from other contexts. On the other hand, if the goal is directly expressed (usually at the beginning of the interaction), the agent needs to understand it. In the following, we briefly note our findings when no goal statement is provided. We then focus more closely on understanding intention from explicitly expressed goal statements.

Use Case 1: without goal statement

Multi-class classification technique is adopted to recognize the user intention. The input are current contexts — 1) time, location; 2) previously launched app; 3) user utterance. Similar to the results in app prediction, last app and user utterance outperform other contexts when predict user intention. Fusing all contexts together yields the best performance. Ranks of individual feature and combinations of features are shown in the right side of Table 1.

Use Case 2: with goal statement

The second use case where user initiates the conversation by a high-level command (e.g., "please organize a meeting for me") is illustrated in Fig 4. Let's take cluster-based intention definition as an example. We first segment the semantic space constructed from past interactions based on user's high-level commands as well as other contexts (red dashed lines in Fig 4). During execution time, the user utters a new command (yellow) which would be mapped to a certain cluster

Table 2. Intention clustering of tasks based on utterances, with typical descriptions.

Cluster	Item Examples (task descriptions supplied by participant)
1	"Picture messaging XXX", "Take picture and send to XXX"
2	"Look up math problems", "Doing physics homework", "Listening to and trying to buy a new song"
3	"Talking with XXX about the step challenge", "Looking at my step count and then talking to XXX about the step challenge"
4	"Playing [game] spiderman", "Allocating memory for spiderman"
5	"Using calculus software", "Purchasing Wolfram Alpha on the play store"
6	"Texting and calling XXX", "Ask XXX if she can talk then call her"
7	"Talking and sharing with group mates", "Emailing and texting group members"

within this semantic space. Thus, we can find past experience of similar nature to the input. By using the information provided by such experience, i.e., how the user previously performed these task with domains, our agent is able to effectively map the new command to a set of supportive domains in the following ways:

- **Representative Sequence** (REPSEQ): We can combine the individual app sequences of the set members into a single app sequence that represents a common way of surfacing the intention. An example is shown in Fig 5. We used ROVER to implement this majority vote [9].
- **Multi-label Classification** (MULTLAB): We can treat this problem as associating multiple labels (app ids) to the input command, given the training instances of cluster members (or neighbors). We used SVM with a linear kernel.

In this use case, we have the following obstacles: 1) people use different language for tasks of similar nature (e.g., "take a picture" vs. "snap a photo"); 2) people use different domains/apps for the same functionality (e.g., GMAIL vs. MESSENGER for contacting someone). The first obstacle holds even for the same user. We adopted the following techniques to solve these problems, with the goal of improving the system's performance at predicting participating apps:

- **Query Enrichment** (QryEnr): We expand the query/ command by incorporating words related to it semantically. QryEnr can reduce the likelihood of seeing sparse input feature vector due to out-of-vocabulary [26] words. The algorithm is shown in Algorithm 1. In short, each word w_i in the lemmatized query q yields mass increases for N semantically close words in the feature vector $\vec{f_q}$.
- **App Similarity** (AppSim): Similarity between two apps is measured in the following ways:
  - Data-driven: App descriptions from the Google Play Store can be projected into a high-dimensional semantic space to compute similarity. We used doc2vec [13] via gensim[1] trained on 1 million app descriptions. Cosine similarity can then be computed given any two apps. Most objects will have associated descriptive materials and we expect this approach can scale accordingly.
  - Knowledge-driven: The Google Play store provides a finite ranked list of "similar apps" for each entry. We used reversed rank $(1/r)$ as similarity.
  - Rule-based: The app package names can be useful, e.g., com.lge.music is close to com.sec.android .app.music since both contain the string "music". Here, we applied a manually constructed list of 50 filters (e.g., "com", "android", "lge") on package names. Then we compute normalized Edit Distance based similarity.

[1] https://radimrehurek.com/gensim/models/doc2vec.html

Algorithm 1 Query Enrichment

Require: lemmatized words of the query $q = \{w_1, ..., w_{|q|}\}$ and their counts $C = \{c_1, ..., c_{|q|}\}$; training vocabulary V; bag-of-words feature vector $\vec{f_q} = \{f_1, ..., f_{|V|}\}$ constructed on q; the word semantic relatedness matrix M; the number of semantically similar words N allowed to be extracted for each word in q;

Ensure: an enriched bag-of-words feature vector $\vec{f_q^*} = \{f_1^*, ..., f_{|V|}^*\}$

```
for each w_i ∈ Q do
    Use M to find N words closest to w_i: V_N = {v_1, ..., v_N} ∈ V;
    for each v_j ∈ V_N do
        f*_j = f_j + M_{i,j} × c_i
    end for
end for
return f*_q;
```

We compare the system-generated apps with the ones users actually launched to compute precision, recall and F_1 score. The results are shown in Table 3. Our main finding is that the original gap between the personalized model and the generic model can be effectively reduced by adopting QryEnr and AppSim techniques, while the personalized model performance (i.e. the upper bound of our agent) can be improved

Table 3. Weighted average F_1 score (%) on test set across 14 participants, using bag-of-word features. Average number of clusters, K_C, in the cluster-based approach is 7.0 ± 1.0 for generic models, and 7.1 ± 1.6 for personalized models. The reported numbers are average performance of 20 K-means clustering results. K_N in the neighbor-based condition is 18.5 ± 0.4 for generic models and 4.9 ± 1.4 for personalized models. AppSim is rule-based.

	REPSEQ		MULTLAB	
	Personal	Generic	Personal	Generic
Cluster (baseline)	42.8	10.5	55.1	24.0
+QryEnr	**44.0**	11.0	**56.1**	27.4
+AppSim	—	14.8	—	29.2
+QryEnr+AppSim	—	**15.4**	—	**38.2**
Neighbor (baseline)	50.8	23.8	51.3	19.1
+QryEnr	**54.9**	26.2	**57.0**	22.9
+AppSim	—	30.7	—	24.7
+QryEnr+AppSim	—	**32.7**	—	**30.3**

as well. This result shows that, after deployment, if the agent keeps observing the user performing tasks, it can learn to assist the user in the future. Even if this is a new user, or out of privacy concern, given insufficient personal data, the agent can leverage the generic model obtained from other users.

We varied ways to compute app similarity. The result is shown in Table 4. As we can see, rule-based approach outperforms the other approaches. The reason could be the coverage issue in the other two methods: since vendor apps do not have entries in our snapshot of Google Store database, 15.5% of the cells of the data-driven similarity matrix are non-zero. For knowledge-driven matrix, only 1.0% are non-zero. Combining three similarity measurements together provides the best performance.

Table 4. Comparison of different AppSim approaches on neighbor-based intention in a generic model. Precision, recall and F_1 score are reported. For the data-driven method, the vector dimension $D = 500$.

	REPSEQ			MULTLAB		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Baseline	33.3	18.9	23.8	45.8	12.3	19.1
Rule	43.3	24.3	30.7	59.4	15.9	24.7
Knowledge	41.8	22.3	28.7	53.0	14.6	22.6
Data	38.1	21.2	27.0	54.6	13.9	21.7
Combine	**44.7**	**25.0**	**31.7**	**61.0**	**16.4**	**25.5**

LANGUAGE REFERENCE

Revealing the agent's understanding of the user's intention can improve the channel transparency. One useful modality is for the agent to verbally convey such understanding (e.g., "I think you want to *plan a trip*"). We adopted keyphrase extraction [2] on user-generated language to generate a ranked list of phrases. We used Rapid Automatic Keyword Extraction (RAKE[2]) algorithm [2], an unsupervised, language- and domain-independent extraction method, reported to outperform other unsupervised method such as TextRank [17, 12] in both precision and F score. In RAKE, we required that 1) each word have 3 or more characters; 2) each phrase have at most 3 words; and 3) each key word appear in the text at least once. We did not tune these parameters. We used 3 individual resources and 2 combinations, reflecting constraints on the availability of different contexts in real-life. The three individual resources are manual transcription of user utterances from their dialogs (MANUAL), ASR transcriptions (ASR) thereof and high-level task descriptions (DESC). The number of phrases generated from different language resources (and their combinations) are shown in Table 5.

[2] `https://www.airpair.com/nlp/keyword-extraction-tutorial`

Table 5. Mean number of phrases generated using different resources. MANUAL: manual transcription of user utterances. ASR: Google speech recognition transcription of user utterances. DESC: user description of the task.

MANUAL	ASR	DESC	DESC + ASR	DESC + MANUAL
20.0	20.3	11.3	29.6	29.1

We selected 6 users to first review their own clusters, by showing them all cluster members with 1) apps used in the member interaction; 2) dialog reproduced; 3) meta-data such as time, day, address, etc. We let them judge whether each individual phrase (the order is randomized) generated by the system summarized all the activities in the cluster (binary judgement). See example in Fig 6.

We found that, on average, users would find an acceptable phrase within top 2 of the list (average Mean Reciprocal Rank $= 0.64$). This demonstrates that, the agent can generate understandable activity references. An ANOVA did not show significant differences between resources. With more sensitive metrics MAP@K[3] (Mean Average Precision at position K) and P@K (Precision at position K) metrics, DESC+ASR and DESC+MANUAL do best. The improvement becomes significant as K increases: having a user-generated task description is very useful.

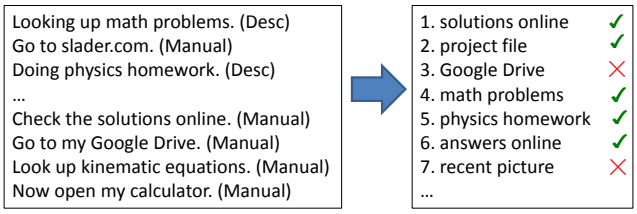


Figure 6. Key phrases (ranked) extracted from user-generated language, with user judgment.

REMAINING CHALLENGES

We have shown that it's possible to build models to infer the user's intention and use this information to activate the set of domains that will assist the user to accomplish the high-level goal. At the same time, we have found that an agent can use the user's language to generate understandable reference to

[3] MAP@K $= \sum_{k=1}^{K} precision(k) * relevance(k)/K$

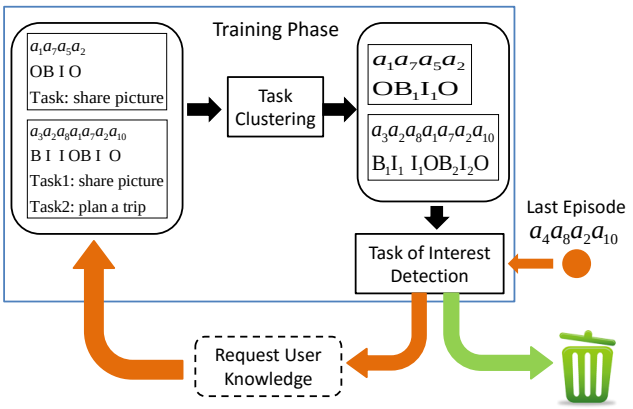


Figure 7. Pipeline for growing complex task inventory

this goal in conversation. Nevertheless there are still remaining challenges to solve before the agent is able to actively help the user in high-level activities.

One challenge is that having to ask the user to explain what they are doing is intrusive and perhaps not realistic. This would be time-consuming in realistic settings. Ideally, the agent should have the ability to balance the cost of an interruption and the expected value of the information to be gained and act accordingly. In practical terms, the agent could observe the user for some period of time to first identify activities that appear to recur. Only when such activities are recognized should the agent ask for a description. We expect that the agent would continue to monitor activities to detect changes or even propose ones itself.

Another challenge is the need for the agent to manage activity-level context so that relevant information can be transferred between apps; for example, passing the address of a restaurant from a reviews app to an app that will provide directions on how to get there. Unless perhaps the apps were developed by the same vendor, its unlikely that similar concepts will be easy to match across different domains. But doing so is necessary for maintaining a context.

Data Gathering

There are explicit and implicit ways to acquire knowledge about the user's high-level intentions so that the inventory of intentions can evolve over time. First, the user could explicitly teach the agent about an activity by saying "To ***plan a trip***, you should find a cheap flight from PRICELINE and then look for 3-star hotel in downtown by using HOTWIRE ..." Such instructable agents have been studied in dialog setup [23, 1]. The difficulties lie in the dependence on the agent's capability to comprehend complex language instructions. Alternately, the user could say "***Agent, watch this***" allowing the agent to observe an activity and then ask the user for more information as needed. This initiative command ("*Agent, watch this*") can be further omitted if the agent is capable of segmenting stream of events into meaningful sessions, e.g., thresholding the idle time.

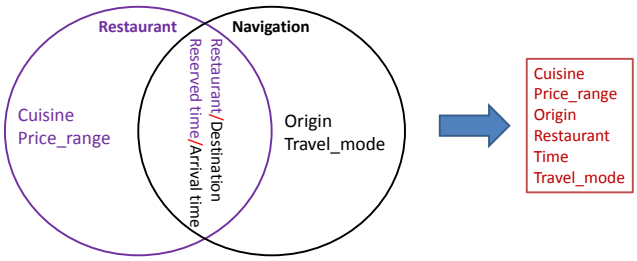


Figure 8. Illustration of overlapping domain knowledge. The size of shared context (red) is less than the sum of concepts in individual domain (purple and black).

A proposed pipeline for this process of knowledge acquisition is shown in Fig 7. We assume that the initial inventory has full annotation from user for each task — 1) apps (e.g., $a_1a_7a_5a_2$ where the indicies indicate the app ids); 2) IOB labels [20] assigned to the apps to distinguish noise (O's) from content (B's and I's); 3) the user's description of what the task is about.

After deployment, the agent first clusters seen interactions (complex tasks in the inventory) into groups and indexes the IOB labels in each interaction accordingly. For example, tasks similar to "sharing picture" are grouped into cluster 1. Then their corresponding IOB labels would be indexed (e.g., $OBIO \rightarrow OB_1I_1O$). Next, the agent detects a task of interests (ToI) — either a recurrence of a seen task type or an anomaly which does not conform to the automatically learned tasks (i.e., normal behavior) [3]. Either way, the system can take appropriate actions to add the current task into the task inventory during the execution time (orange path in Fig 7) as elaborated in Table 6.

Sharing Context

The necessary communication skills to request concepts or slots in individual domains are already provided by domain experts (e.g., "`request_destination`" in NAVIGATION domain, "`inform_restaurant_review`", "`request_price_range`" in RESTAURANT domain). It is up to the agent to mix these skills into a unified conversation. However, the concepts required by different domains may overlap. For example, `destination` in NAVIGATION domain may be implied by the `restaurant` in RESTAURANT domain. More intelligently, the `arrival_time` in NAVIGATION can be computed as `reserved_time` - 10min by understanding that user tends to arrive 10min earlier. Therefore, it is important for the agent to have a shared context in order to avoid requesting information it has already possessed.

An illustration is shown in Fig 8. Two domains (RESTAURANT and NAVIGATION) have a few concepts in common but with different names. When collectively serving a common user intention ("plan a dinner"), knowing the value of `Restaurant` slot induces the value of `Destination`. Thus, this problem can be formalized as follows: Given the union of concepts in D different domains $C = C_1 \cup C_2 \cup ... \cup C_D$ and the concepts already filled in $C_F \in$

Table 6. System actions to grow task inventory based on ToI type classification and confidence

ToI Type	Confidence	System Action	Example System Prompts
Recurrence	High	Add to seen interactions	N/A
Recurrence	Mild	Confirm with user	"I think you were *planning a party*, am I right?"
Recurrence	Low	Request user annotation	"Please tell me what you just did. Is it one of `[list of tasks]`?"
Anomaly	N/A	Request user annotation	"I think you were doing something new, could you teach me?"

C, for the target concept c_t find the source concept $c_s = \arg\max_{c_i \in C_F} R(c_i, c_t|\text{Intention})$, where function R measures the relatedness (semantic similarity) between two concepts given the ultimate user intention. A further filtering function has to be applied to either use the source concept c_s (e.g., `restaurant`) as the target concept (`destination`) or discard it. A perfect R measurement plus the filtering function would resolve the inter-domain as well as intra-domain redundancy.

The following approaches can be adopted to learn the semantic relatedness function $R(c_s, c_t|\text{Intention})$ where we assume c_s and c_t are from different domains:

1. Rule-based: In a multi-domain conversation, if the values of c_s and c_t coincide, they are probably of similar nature;
2. Data-driven: We can embed concepts C in the training dialog corpus. Thus, c_s and c_t can be projected to a semantic space and their similarity can be computed.

We believe that this class of information could be pooled across users: identifying the right mappings in principle needs to be done only once for any given pair of apps, with extensions being inferred through transitivity. At this point, this is speculative. But we believe that it can be part of a strategy for establishing an operational ontology across apps.

CONCLUSION

We present a framework that will allow an agent to implicitly learn from past interactions to map high-level expressions of goals (e.g., "go out with friends") to specific functionalities (apps) available in a smart environment. The proposed agent uses language produced by the user to identify interactions similar to the current input. A set of participating domains/apps can be proposed from past experience and used to support current activities. This framework is also capable of generating natural language references to past experience clusters. As a result, the communication channel would have greater transparency, supporting timely recovery from possible misunderstandings. The value of such an agent is that it can learn to manage activities at a level more abstract than provided by object-specific interfaces and would allow users to build their own (virtual) apps that combine the functionalities of existing objects.

ACKNOWLEDGMENTS

This work is partially funded by Yahoo! InMind project and General Motors Advanced Technical Center. We thank Zhenhao Hua for implementing the logger app. We thank Yulian Tamres-Rudnicky and Arnab Dash for collecting the data.

REFERENCES

1. Azaria, A., Krishnamurthy, J., and Mitchell, T. M. Instructable intelligent personal agent. In *Proc. The 30th AAAI Conference on Artificial Intelligence (AAAI)* (2016).
2. Berry, M. W., and Kogan., J. *Text mining: applications and theory*. Wiley, 2010.
3. Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. In *ACM computing surveys (CSUR)* (2009).
4. Chen, Y.-N., and Rudnicky, A. I. Dynamically supporting unexplored domains in conversational interactions by enriching semantics with neural word embeddings. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, IEEE (2014), 590–595.
5. Chen, Y.-N., Sun, M., and Rudnicky, A. I. Leveraging behavioral patterns of mobile applications for personalized spoken language understanding. In *Proc. The 18th ACM International Conference on Multimodal Interaction (ICMI)* (2015), 83–86.
6. Chen, Y.-N., Sun, M., Rudnicky, A. I., and Gershman, A. Unsupervised user intent modeling by feature-enriched matrix factorization. In *Proc. The 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016).
7. Chen, Y.-N., Wang, W. Y., Gershman, A., and Rudnicky, A. I. Matrix factorization with knowledge graph propagation for unsupervised spoken language understanding. In *Proc. The 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, ACL (2015).
8. Duda, R., Hart, P., and Stork, D. *Pattern Classification*. John Wiley and Sons, 2012.
9. Fiscus, J. G. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)* (1997), 347–352.
10. Harrison, C., Xiao, R., Schwarz, J., and Hudson, S. E. Touchtools: leveraging familiarity and skill with physical tools to augment touch interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 2913–2916.

11. Hastie, H., Aufaure, M.-A., Alexopoulos, P., Bouchard, H., Breslin, C., Cuayhuitl, H., Dethlefs, N., Gaic, M., Henderson, J., Lemon, O., Liu, X., Mika, P., Mustapha, N. B., Potter, T., Rieser, V., Thomson, B., Tsiakoulis, P., Vanrompay, Y., Villazon-Terrazas, B., Yazdani, M., Young, S., and Yu, Y. The Parlance mobile application for interactive search in english and mandarin. In *Proc. The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, ACL (2014), 260–262.

12. Hulth, A. Improved automatic keyword extraction given more linguistic knowledge. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL (2003), 216–223.

13. Le, Q. V., and Mikolov, T. Distributed representations of sentences and documents. In *Proc. The 31st International Conference on Machine Learning (ICML)* (2014), 1188–1196.

14. Li, Q., Tür, G., Hakkani-Tür, D., Li, X., Paek, T., Gunawardana, A., and Quirk, C. Distributed open-domain conversational understanding framework with domain independent extractors. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, IEEE (2014), 566–571.

15. Lin, B.-s., Wang, H.-m., and Lee, L.-s. A distributed architecture for cooperative spoken dialogue agents with coherent dialogue state and history. In *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, vol. 99 (1999), 4.

16. Lunati, J.-M., and Rudnicky, A. I. Spoken language interfaces: The OM system. In *Proc. Conference on Human Factors in Computing Systems* (1991), 453–454.

17. Mihalcea, R., and Tarau, P. TextRank: Bringing order into texts. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL (2004), 404–411.

18. Nakano, M., Sato, S., Komatani, K., Matsuyama, K., Funakoshi, K., and Okuno, H. G. A two-stage domain selection framework for extensible multi-domain spoken dialogue systems. In *Proc. The 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, ACL (2011), 18–29.

19. Pappu, A., Sun, M., Sridharan, S., and Rudnicky, A. I. Situated multiparty interaction between humans and agents. In *Human-Computer Interaction* (2013).

20. Ramshaw, L. A., and Marcus, M. P. Text chunking using transformation-based learning. In *Proc. ACL Workshop on Very Large Corpora* (1995).

21. Raux, A., Langner, B., Black, A. W., and Eskenazi, M. LET'S GO: Improving spoken dialog systems for the elderly and non-native. In *Proc. The 8th European Conference on Speech Communication and Technology (Eurospeech)* (2003).

22. Rudnicky, A. I., Lunati, J.-M., and Franz, A. M. Spoken language recognition in an office management domain. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE (1991), 829–832.

23. Rudnicky, A. I., Pappu, A., Li, P., Marge, M., and Frisch, B. Instruction taking in the teamtalk system. In *AAAI Fall Symposium: Dialog with Robots* (2010), 173–174.

24. Ryu, S., Song, J., Koo, S., Kwon, S., and Lee, G. G. Detecting multiple domains from users utterance in spoken dialog system. In *Proc. The 6th International Workshop on Spoken Dialogue Systems (IWSDS)* (2015), 101–111.

25. Sun, M., Chen, Y.-N., Hua, Z., Tamres-Rudnicky, Y., Dash, A., and Rudnicky, A. I. Appdialogue: Multi-app dialogues for intelligent assistants. In *Proc. International Conference on Language Resources and Evaluation (LREC)* (2016).

26. Sun, M., Chen, Y.-N., and Rudnicky., A. I. Learning OOV through semantic relatedness in spoken dialog systems. In *Proc. The 16th Annual Conference of the International Speech Communication Association (Interspeech)* (2015), 1453–1457.

27. Sun, M., Chen, Y.-N., and Rudnicky, A. I. Understanding user's cross-domain intentions in spoken dialog systems. In *NIPS workshop on Machine Learning for SLU and Interaction* (2015).

28. Sun, M., Chen, Y.-N., and Rudnicky, A. I. HELPR a framework to break the barrier across domains in spoken dialog systems. In *Proc. The 7th International Workshop on Spoken Dialog Systems (IWSDS)* (2016).

29. Sun, M., Chen, Y.-N., and Rudnicky, A. I. An intelligent assistant for high-level task understanding. In *Proc. The ACM Conference on Intelligent User Interfaces (IUI)* (2016).

30. Tibshirani, R., Walther, G., and Hastie., T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63* (2001), 411–423.

31. Young, S. Using POMDPs for dialog management. In *Proc. IEEE Spoken Language Technology Workshop (SLT)* (2006), 8–13.

A Proposal for Qualities for Smart Objects

James L. Crowley
Université Grenoble Alpes
Laboratoire Informatique de Grenoble
INRIA Grenoble Rhônes Alpes Research Center
James.Crowley@inria.fr

Joelle Coutaz
Université Grenoble Alpes
Laboratoire Informatique de Grenoble
Joelle.Coutaz@imag.fr

ABSTRACT
In this position paper, we explore qualities for describing the autonomy, suitability and composability of smart objects. The specific set of qualities proposed below is tentative and incomplete. Our intention is to trigger discussion and debate within the scientific community in order to arrive at a consensus. If successful, a longer, more complete paper will then be prepared to communicate the results.

Author Keywords
Software Qualities, Smart Objects, Situated Interaction

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

SMART OBJECTS
Smart objects are ordinary objects that have been augmented with computation and communication as well as abilities for perception, action and/or interaction. Smart objects can work individually or collectively to provide services for users where services can be defined as actions and/or information that provide value. As the field of smart objects matures into a recognized engineering discipline, there is an increasing need for tools to model and predict the properties of smart objects. Qualities provide a principled approach to define metrics for performance and are essential for a scientific approach to the development of smart objects.

QUALITIES
Qualities are characteristics that can be used as normative references for the development process of products, from requirements specification to evaluation, as well as for products comparison.

As global trade has matured, the need for international agreements on quality assessment has become important. It is not surprising then that we are facing the existence of a number of quality models from distinct areas. What these models have in common is the hierarchical organization of quality characteristics into sub-characteristics where the leaves consist of quality attributes that can be measured.

Many terminologies and concepts are used for discussing qualities in different fields, and it is not uncommon to find different terms for similar concepts or even to find the same term used with different meanings within different communities. It is also not uncommon to find terms and concepts at different levels in a hierarchical organization, depending on the nature of the objects studied within different communities. What is most important is that members within a community reach agreement on the concepts and terms for quality as well as their hierarchical relations.

Smart objects are hybrid systems composed of physical and software elements. Thus it is natural to use qualities from both manufacturing and software engineering as inspiration for qualities for smart objects. Garvin proposed eight critical dimensions of quality that serve as a framework for analysis in manufacturing: performance, features, reliability, conformance, durability, serviceability, aesthetics, and perceived quality [1].

Since the seminal work of McCall [2] and Boehm [3], the software engineering community has developed a variety of quality models until some consensus was reached with the ISO/IEC 9126 proposition [4]. To reflect the different stages in the software development process, a distinction is made between quality in use, external quality and internal quality. *Quality in use* is "the capability of the software product to enable specified users to achieve specified goals with *effectiveness*, *productivity*, *safety* and *satisfaction* in specified contexts of use".

Quality in use contributes to the specification of the requirements for (and depends on) external quality. *External quality* is "the quality when the software end product is executed, which is typically measured and evaluated while testing in a simulated environment with simulated data". In turn, external quality defines the requirements for (and depends on) the *internal quality* (i.e. the developer's view of the system under development). External quality is structured into six characteristics: functionality, reliability, usability, efficiency, maintainability, and portability. These categories are refined into sub-categories. For example, usability covers understandability, learnability, operability, and attractiveness.

IUI 2016 Workshop: Interacting with Smart Objects, March 10th, 2016, Sonoma, CA, USA

In Human Computer Interaction, The IFIP WG 2.7(13.4) on Engineering HCI has concentrated on the understandability and operability aspects of usability, but with the use of different terms [5]: *interaction flexibility* to refer to "the multiplicity of ways the user and the system exchange information during task execution" and *interaction robustness* to denote "the capacity of the system to support users in achieving their goals successfully". These subcategories are in turn refined into measurable properties.

All these early (and ongoing) works on quality models and quality assessment are motivated by rationalism, utilitarism and immediacy. There is very little room left to phenomena that develop over time such as attachment, or to characteristics of products that exceed expectation, nourish dreams, or simply human values.

Based on these early works, we propose a three-level hierarchy of qualities. At the top level, we promote four families of qualities that concern Autonomy, Suitability, Durability and Composability. Within these families, smart objects can be described by intrinsic and extrinsic qualities.

Intrinsic qualities [6] are properties that describe a smart object, independent of any interaction with its operating environment. Intrinsic qualities include such characteristics as size, weight, and power requirements as well as properties such as reliability, availability and security. In most cases, intrinsic qualities reflect how well a system or object complies with functional requirements. These can be used to define performance metrics that can be tested under controlled laboratory conditions.

Extrinsic qualities describe how the object or system interacts with its external environment, including users. We include durability, usability, controllability and trustworthiness. Measurement of extrinsic qualities generally requires deployment and evaluation under real world conditions.

Time emerges as an interesting base for organization. In general, for most qualities, there is a basic static definition, and a set of possible projections over time. When possible, we will follow this model for defining qualities.

AUTONOMY

Autonomy is the ability of a system to maintain its own integrity. For a smart object, autonomy depends on a number of component qualities including energy management, reliability, and durability. Thus autonomy is a composite quality whose definition and measurement depends on it components.

Failure, for a smart object, is the partial or total loss of system availability. Failure can be temporary or permanent and can refer to specific functionality or overall system operation. Many different quality metrics refer to conditions of failure.

Power autonomy

Systems require energy and consume power. Energy is the capacity of a system to perform work, measured in Joules. Power is a measure of the energy consumed per unit time, measured in Joules/Second. Average power consumption, as well as minimum and maximum instantaneous power requirements are important qualities for any system.

Smart objects typically rely on electrical power for energy. Electrical power is measured in Watts defined as Joules/Second. Minimum and maximum power requirements of a smart object are measured in Watts, while average power consumption is measured in Watt-Hours. Standard tools and techniques exist for measuring electrical power consumption, and these can and should be adopted for measuring and monitoring power consumption.

Power autonomy is the ability to operate without a physical connection to an external power supply. The need for power autonomy frequently raises important challenges in the design of smart objects. In addition to instantaneous and average power requirements, the duration of power autonomy can be an important factor in the commercial value of smart objects.

Reliability

Reliability is the ability of a system to consistently perform its required function without degradation or failure over time.

Among the most common measures of reliability are the mean time to first failure, the mean time between failures (MTBF) and the failure rate per unit time [1] where failure can be a partial or total loss of correct function or even a loss of availability.

Availability can be an important component of reliability. Availability measures "the proportion of total time during which the system is an up state" [2], that is how often the object is available for use, even though it may not be functioning correctly. Availability can enable recovery through external intervention.

More detailed measures of reliability depend on specific functions. These can include probability of error, and failure rate.

Durability

Durability is the ability to withstand repeated use over a period of time without significant deterioration in performance. Durability is distinct from reliability in that it includes the way the object is used as well as the time over which it operates [1]. Durability can be defined as the amount of use one gets from an object before it fails while reliability is the length of time between failures regardless of use. For example, the number of recharge cycles of a lithium ion battery is a measure of its durability, rather than its reliability.

For smart objects, two notions of durability are relevant. The first, obvious, notion refers to the device life cycle

(longevity). This can be measured by the number of hours of operation or by the number of times a device can be turned on and off.

A more interesting notion of durability captures the ease with which a device resists obsolescence. Causes for obsolescence are three-fold: software, material, and people. For the software components of a smart object, durability can be measured in terms of maintainability and portability. For the material/hardware components, durability can be measured with the lifespan of its materiality, but also by its capacity to be re-cycled and up-cycled, along with its environmental impact from the extraction of raw material to disposal.

People discard some objects, although they are still functioning, while they preserve others because of some form of bonding that builds over time. Early studies show that factors that affect attachment include [7]: engagement ("the extent to which an object invites and promotes physical engagement with its owner during use"), histories ("the extent to which the object preserves personal memories"), and augmentation ("the extent to which an object can be reused, modified, altered beyond its original use". Composability, which permits incremental changes, is thus an important characteristics to consider for smart objects.

COMPOSABILITY

Composability refers to the ability of the object to function as part of an assembly to collectively provide a set of services that makes sense for users. Objects can be assembled mechanically, electrically, wirelessly or functionally. The essential features that make a smart object amenable to composition are the generality, diversity, cardinality and difficulty of its interconnections.

Generality, diversity, cardinality and difficulty can be illustrated using electrical interconnection. An object can be characterized by the number of different types of physical connectors (diversity), the number of different types of other objects that can be connected via its physical connectors (generality), and the total number of physical connectors available (cardinality). Similarly, for wireless interconnection, one can describe the generality of the wireless protocols available, the diversity of the types of wireless protocols, and the cardinality of connections that can be maintained at any one time.

Functional composition can be formalized in terms of software services [8]. An object can be characterized by the overall number of composite services that can make use of its functions (generality), the variety of different kinds of composite services to which it can contribute (diversity), and the number of composite services to which it can contribute at the same time (cardinality). Pushing the analysis further, a distinction can be made between syntactic composability (e.g., are data types and parameter passing compatible?), and semantic composability (e.g., is data from the source service within bounds expected by the sink service?).

Additional aspects of composability concern property preservation, temporal aspects of composition, and the amount of effort required to establish an interconnection:

- Does the assembly preserve the properties carried by the smart objects individually? For example, if two smart objects satisfy observability, is observability still supported by the composition (in particular, is the state of the interconnexion observable)?
- Is the composition - reconfiguration, and decomposition, static or dynamic? In turn, dynamicity supposes the existence of the appropriate underlying middleware whose functioning should be amenable to people.
- Is the assembly performed automatically by the objects themselves, manually by users, or by the cooperation of both? In case of human intervention, what are the cognitive-sensory-motor efforts required? What is the most suitable approach?

SUITABILITY

Suitability is the capacity to act and interact in a manner that is appropriate for the task and context. Suitability can be characterized by a variety of properties ranging from intelligence to usability, controllability, aesthetics and trust.

Intelligence

Simply augmenting an object with micro-electronics does not make it smart. To be smart, an object should behave in a manner that is appropriate for its role and its environment. Indeed, in many cases the most basic quality for a smart object is smartness, defined as the appropriateness, or intelligence, of its behaviors as judged by an external observer. In robotics, appropriate behavior is commonly referred to as situated [9].

Usability

As discussed above, a number of quality models, factors and criteria have been developed to define and assess the usability of interactive systems from desktop computers, tablets, tabletops, smart phones, and even cars and mobile robots. All of these are valid for assessing the usability of smart objects, although the notion of "useworthiness" is not sufficiently addressed.

Useworthiness is central to Cockton's argument for the development of systems that have value in the real world [10]. In value-centered approaches, design starts with an explicit expression of the potential values for a set of target contexts. Intended value for target contexts are then translated into evaluation criteria. Evaluation criteria are not necessarily elicited from generic intrinsic features such as time for task completion, but are contextualized. They are monitored and measured in real usage to assess the achieved value.

Controllability

For smart objects, controllability is the ability to regulate, dominate or command the behaviors (action and interactions) of the object. Control of ones' personal environment is an important component of general well being, and can be an important factor in the rate with which individuals will invest time and money in smart objects and smart services.

Loss of control (or preemption) is an important aspect to measuring controllability. Thus, controllability can be measured in terms of the number of behaviors that can or cannot be selected at any time,

Privacy and Security

Privacy is the ability to protect information from disclosure or observation. Security is the ability to assure that both information and system components have not been subject to unauthorized modification or disclosure. Privacy and security are closely related but separate properties.

Note that security is distinct from trust. Security refers to the ability of the system to withstand attack, while trust refers to the confidence that users have in the ability of the system to withstand attack. As system may be secure but untrusted, or it may be trusted but insecure.

As a form of protection, both privacy and security are measured as absence of violations. For privacy, this means explicitly listing all information that is or can be disclosed. For security, this means listing the category of attacks that the system is certified to resist.

Trustability

Trustability is ability to inspire confidence in one or more qualities. This can range confidence in the reliability of the object, in the intelligence of the object or in its privacy and security. As mentioned above, trust in the quality of an object is separate from the quality itself. Because trust is an ability to inspire users, the obvious manner to measure trust is to gather statistics on the beliefs of users with regard to different qualities.

CONCLUSIONS

In this short position paper we have explored definitions for qualities that describe the Autonomy, Suitability and Composability of smart objects. For each family, we have defined intrinsic qualities that concern innate properties of a smart object as extrinsic properties that describe the interaction between the object and its environment. We have noted that many qualities have a core static definition as well a set of possible temporal projections.

The definitions presented above are tentative and incomplete. Our purpose is to trigger discussion and debate in order to arrive at a true consensus within the scientific community. If successful, the community will have made an important advance towards establishment of the study of Smart Objects as a scientific discipline.

Many of the desired qualities of smart objects arise from a goal of improving quality of life. Quality of life (QoL) is commonly defined as the general well-being of individuals and societies. Metrics for quality of life have been proposed for fields such as healthcare and gerontology. These can be used to define qualities for smart object based on the requirements of target application domains.

The ultimate quality for any system is value. The obvious measure of value is how much wealth a user is willing to pay to assure access to the system. However, there are aspects to value that in some cases defy monetary estimates. Measuring value requires understanding the services that a system provides to a user, and how much sacrifice the user is willing to undergo to assure access to the service.

ACKNOWLEDGMENTS

This work and ideas reported in this paper have been partially sponsored by the French Agence Nationale de la Recherche (ANR), program "Investissement d'Avenir" project reference ANR-11-EQPX-0002, Amiqual4Home and European program CATRENE project AppsGate (CA110).

REFERENCES

1. D. A. Garvin, Competing on the Eight Dimensions of Quality, Harvard Business Review, November 1987.
2. J. A. McCall, P. K. Richards and G. F. Walters, Factors in software quality. General Electric, National Technical Information Service, 1977.
3. B. Boehm, J. R. Brown, H Kaspar, M. Lipow, G. J. MacLeod, M. J. Merritt, Characteristics of Software Quality, North-Holland, 1978.
4. ISO/IEC 9126-1:2001, "Software engineering – product quality- Part1: Quality Model, 2001.
5. C. Gram, G. Cockton, "Design Principles for Interactive Software", Chapman & Hall, 1996.
6. B. B. Agarwal, S. P. Tayal, M. Gupta, Software Engineering and Testing, Jones and Bartlett, 1975.
7. M. Odom, J. Pierce, E. Stolterman, and E. Blevis, "Understanding Why We Preserve Some Things and Discard Others in the Context of Interaction Design", Proc. CHI'09, ACM, pp. 1053-1062.
8. J. L. Crowley, J. Coutaz, "An Ecological View of Smart Home Technologies", 2015 European Conference on Ambient Intelligence, Athens, Greece, Nov. 2015.
9. C. Breazeal, Designing Sociable Robots, MIT Press, Cambridge MA, 2002.
10. G. Cockton, "A Development Framework for Value-Centred Design", CHI'05 extended abstracts on Human factors in computing systems, pp. 1292-1295, ACM, 2005.

NLU vs. Dialog Management: To Whom am I Speaking?

Dirk Schnelle-Walka
Harman International
Germany
dirk.schnelle-walka@harman.com

Stefan Radomski
TU Darmstadt
Germany
radomski@tk.informatik.tu-darmstadt.de

Benjamin Milde
TU Darmstadt
Germany
milde@lt.informatik.tu-darmstadt.de

Chris Biemann
TU Darmstadt
Germany
biem@lt.informatik.tu-darmstadt.de

Max Mühlhäuser
TU Darmstadt
Germany
max@informatik.tu-darmstadt.de

ABSTRACT
Research in dialog management and natural language understanding are both approaching voice-based interaction. Coming from different perspectives they emphasize different components in the spoken dialog system processing chain. Although each approach is suitable to provide a satisfiable user experience, a combined approach could potentially improve towards a more convincing natural interaction with the user as discussed in this vision paper.

Author Keywords
natural language understanding; dialog management; intelligent personal assistants; user experience

ACM Classification Keywords
H.5.m User Interfaces: Miscellaneous

DIALOG MANAGEMENT
Speech is considered to provide an efficient and pleasant way to interact with smart objects [29]. Historically, these systems were built along a processing chain to actually initiate actions based on the user's utterance and/or produce spoken output in return. A general architecture, according to Kunzmann [13], of these system is shown in Figure 1. Pieraccini describes the components in [21] as follows: The Automated Speech Recognition (ASR) component converts the raw audio input into a sequence of words (or the n-best results). This is forwarded to a Natural Language Understanding (NLU) component to extract the semantics of the utterance. This is used by the dialog manager (DM) to decide upon the action to take according to the employed dialog strategy. The DM may use stored contextual information derived from previous dialog turns. One of the actions, a DM may take is the generation of spoken output. Therefore, the response generation (RG) generates text as an output which is passed to the text-to-speech engine (TTS) component to be synthesized into an utterance.

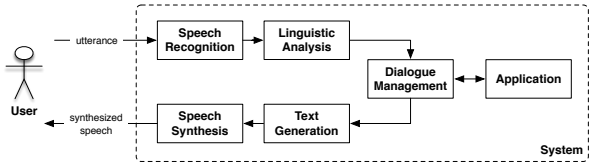


Figure 1. General architecture of a spoken dialog system

Research has been centered around DM for many years. One of the main efforts was the development of suitable dialog strategies for a more natural user experience. Radomski provides a thorough analysis of the related terms in [22] for *dialog*, *dialog management* and *dialog strategy*. Based on various definitions throughout literature, e.g. by Traum [27] or Rudnicky [23] he comes the following definitions for multimodal dialogs. We adapted them to voice user interfaces.

Definition 1 *A **dialog** is a sequence of interleaved, communicative events between a human and a computer to convey information aurally.*

Definition 2 *A **Dialog Manager** is a software component responsible for maintaining the dialogs state and driving the interaction by mapping relevant user input events onto system responses as output events. Performing these responsibilities is also referred to as dialog management.*

IUI 2016 Workshop: A joint Workshop on Smart Connected and Wearable Things, March 10th, 2016, Sonoma, CA, USA

DOI: 10.13140/RG.2.1.1928.4247
urn:nbn:de:tuda-tuprints-54208

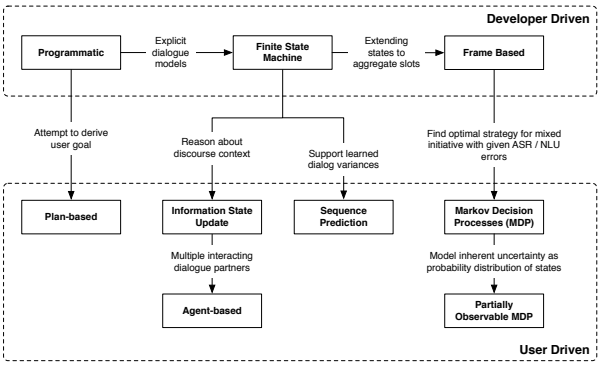


Figure 2. Pattern Language for Dialog Management

Definition 3 *A **Dialog Strategy** is a conceptualization of a dialog for an operationalization in a computer system. It defines the representation of the dialogs state and respective operations to process and generate events relevant to the interaction.*

Schnelle-Walka et al. [24, 25] developed a pattern language thereof as shown in Figure 2. They identified the following strategies: (i) Programmatic Dialog Management, (ii) Finite State Dialog Management, (iii) Frame Based Dialog Management, (iv) Information State Update [14], (v) Plan Based, (vi) Markov Decision Process [16] and (vii) Partially Observable MDP [31]. Each strategy has its strengths and weaknesses. Some are more restricted while others allow for less constrained user input. Generally, the system used to define the degree of freedom that users have while interacting with the system. They all share the DM-centered perspective regarding the NLU to be some input into the system while the decision upon subsequent interaction is being handled in this component.

This concept has also been applied to multimodal approaches to DM, like PAC-AMODEUS [6], TrindiKit [15], Jaspis [28] or MIMUS [2] as well as high level architectures [17].

NATURAL LANGUAGE UNDERSTANDING

Natural language understanding (NLU) is a subtopic of natural language processing in artificial intelligence that deals with machine reading [8] comprehension. NLU targets the automatic comprehension of entire documents without anticipating their content.

In the past years, performance of NLU increased dramatically as sketched by Cambria [5] and shown in Figure 3. Today's NLU moves away from the *Syntactical Curve* to the *Semantic Curve*. While the previous one focuses on processing of documents at a syntax level, like keyword or word co-occurrence count, newer concepts "rely on implicit denotative features associated with natural language text" [5].

Research in NLU usually learns from large document sets. One application that demonstrates the level of understanding is to query for information, also known as *Question Answering*. This is, where interaction with the user comes into play.

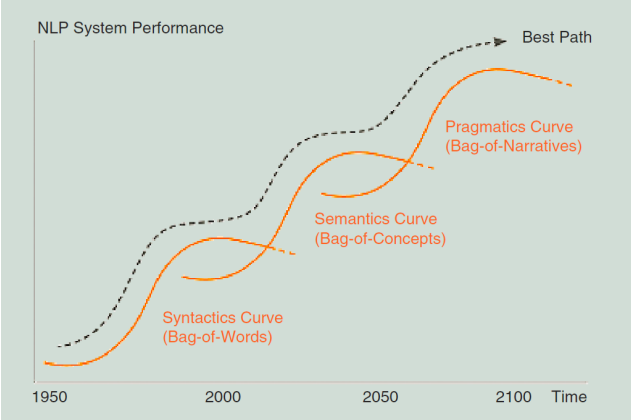


Figure 3. Envisioned evolution of NLU research through different eras or curves

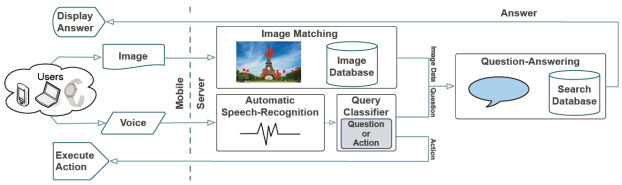


Figure 4. Pipeline of Sirius as an example for an intelligent personal assistant

Hence, a typical example is seen in the development of intelligent personal assistants (IPA). One example of such an IPA is the open source IPA Sirius from Hauswald et al. [10] as shown in Figure 4. Other examples of IPAs that expose their API to developers include IBM Watson [9] and LUIS [1] from Microsoft. While the first IPAs were only able to cope with a single dialog turn, newer systems also establish dialog context. Thus, they are able to also refer to previously entered input and, e.g. iteratively refine query results by adding or removing parameters as needed. This way, they are adopting tasks, such as maintaining the conversational state, that researchers in dialog management see as one of the core tasks of a DM.

CONTRASTING NLU AND DM

While the AI community usually focuses on NLU, the spoken dialog community focuses on the DM as the central point in this chain. Both have good reasons for their approach and are able to deliver convincing results.

DM-centered systems are principally constrained because they anticipate the users input as plans to help them to achieve their goal. Depending on the implemented dialog strategy they allow for different degrees of flexibility.

NLU-centered systems see the central point in the semantics of the utterance, which should also be grounded with previous utterances or external content. Thus, whether speech or not, NLU regards this as a stream of some input to produce some output. Since no dialog model is employed, resulting user interfaces currently do not handle much more than single queries.

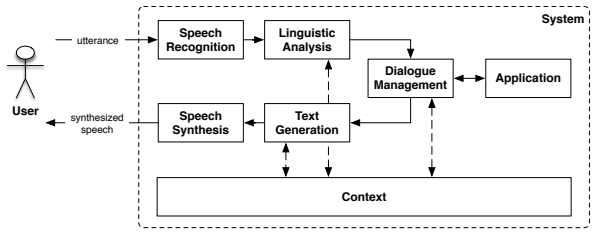


Figure 5. General architecture of a spoken dialog system including context management

Currently, efforts towards spoken interaction coming from this domain are still not fully aware of what has been done in DM research in the past decades, and vice versa. Both parties are coming from different regions in the chain of spoken dialog systems. For instance, `api.ai` recently announced that their system now supports slot filling[1]. The biggest challenges are seen in determining the user's intent and semantic slot filling [18]. The user may use these to refine a query until he ends with a single result. Current spoken dialog systems are already beyond that and are able provide good voice user interface design. For instance, grounding strategies as they are introduced e.g., by Larsson in his Information State Update approach [14], are not exploited. Another important aspect are dialog acts [3]. Interaction with smart objects must go beyond the question-answer paradigm and rely on, e.g., reject, accept, request-suggest, give-reason, confirm, clarify. And finally, uncertainty in the recognition result [26] is not considered at all. NLU focused systems rely on their ability that the user can replace any value at any time. Therefore, he will have to understand the received result and correct it as needed. The developed strategies for error correction and error prevention, as they have been researched in the DM community for years [4], like explicit or implicit confirmation, remain unexploited.

As it comes to maintaining the conversational state, both NLU and DM will need to access it. NLU will need it, e.g., to correctly determine linguistic phenomena like ellipsis or anaphoric references. DM will need it to allow for a more natural dialog flow to produce the right output following dialog theoretical aspects. Context must be accessible and manipulatable from both components, as shown in Figure 5. This aspect was already addressed, e.g. by Oviat [19] who added a *Context Management* component to the processing chain. Coming from multimodal fusion she demands for a canonical meaning representation.

SUMMARY AND OUTLOOK

In this paper, we had a look at the approaches of the community of DM and the community of NLU to voice-based interaction. We described both views onto it, that emphasize different components in the processing pipeline. Subsequently, we explored synergy effects of both views.

For a more convincing user experience both communities will be in the need of adopting techniques from the other community. The capabilities of today's NLU are already convincing. There are lacks in how to engage the user into a real conversation. These techniques have been well developed in the domain of dialog management. Adoption of dialog theory will allow for a more natural interaction.

We believe, that is time that both communities start talking to each other to better incorporate results of "the other component" to arrive at a convincing user experience. Maybe, POMPDP [31] dialog systems are a good candidate to be employed as they are also based on machine learning techniques that provided a breakthrough in NLU and are the most advanced dialog strategy. Maluuba[2], a Candaian NLU centered company already started rolling out such systems.

However, future systems may differ from what has been described above. *Cognitive Computing* is about to change the way how voice-based interactive systems will be developed in the future. We follow the definition given in [12].

Definition 4 ***Cognitive Computing*** *refers to systems that learn at scale, reason with purpose and interact with humans naturally. Rather than being explicitly programmed, they learn and reason from their interactions with us and from their experiences with their environment.*

This has implications for voice-based interaction: (i) It would be desirable if voice-based system would learn and get better while being used, instead of being statically defined or trained. This can apply to speech recognition, NLU and text generation components, with online learning from implicit or explicit user feedback. Some headway is also being made in the machine learning community in the form of proactive learning [7], as user feedback can be subjective and must be judged according to its information value. (ii) Making voice-based interaction more natural would also entail that responses are not programmed, but produced by a generative model. (iii) The ability to transfer and use knowledge from known domains and tasks to previously unknown and new tasks is also a building block of cognitive computing systems. Dialog systems could also benefit from the transfer learning paradigm [20], as it offers solutions for data scarcity in a particular domain. An example would be a tourist information dialog system that transfers what has been learned in a restaurant recommendation dialog system.

The Recurrent Neural Network (RNN) framework is a candidate that could make (i)-(iii) possible, with some recent and promising first results [11, 30].

REFERENCES

1. Fast and easy language understanding for dialog systems with Microsoft Language Understanding Intelligent Service (LUIS). In *Special Interest Group on Discourse and Dialogue.* (2015), 159–161.
2. Amores, J. G., Pérez, G., and Manchón, P. MIMUS: a multimodal and multilingual dialogue system for the

[1] `https://api.ai/blog/2015/11/09/SlotFilling/`

[2] `http://maluuba.com`

home domain. In *ACL*, Association for Computational Linguistics (2007), 1–4.

3. Austin, J. L. *How to do things with words*, vol. 367. Oxford University Press, 1975.
4. Bohus, D., and Rudnicky, A. Sorry, I Didn't Catch That! - An Investigation of Non-understanding Errors and Recovery Strategies. In *SIGdial Workshop on Discourse and Dialogue* (2005).
5. Cambria, E., and White, B. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine* (2014), 48–57.
6. Coutaz, J. PAC: An object-oriented model for dialog design. In *Human Computer Interaction (INTERACT)* (1987), 431–436.
7. Donmez, P., and Carbonell, J. G. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *ACM conference on Information and knowledge management* (2008), 619–628.
8. Etzioni, O., Banko, M., and Cafarella, M. J. Machine Reading. *AAAI 6* (2006), 1517–1519.
9. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. a., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., and Welty, C. Building Watson: An Overview of the DeepQA Project. *AI Magazine 31*, 3 (2010), 59–79.
10. Hauswald, J., Laurenzano, M. A., Zhang, Y., Li, C., Rovinski, A., Khurana, A., Dreslinski, R. G., Mudge, T., Petrucci, V., Tang, L., and Mars, J. Sirius : An Open End-to-End Voice and Vision Personal Assistant and Its Implications for Future Warehouse Scale Computers. In *Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)* (2015).
11. Henderson, M., Thomson, B., and Young, S. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Spoken Language Technology Workshop (SLT)*, IEEE (2014), 360–365.
12. Kelly, J. E. Computing, cognition and the future of knowing. Whitepaper, IBM Reseach, 2015.
13. Kunzmann, S. Applied speech processing technologies-our journey. *European Language Resources Association Newsletter (ELRA)* (2000), 6–8.
14. Larsson, S. *Issue-based Dialogue Management*. PhD Thesis, University of Gothenburg, 2002.
15. Larsson, S., and Traum, D. R. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering 6*, 3&4 (2000), 323–340.
16. Levin, E., Pieraccini, R., and Eckert, W. Using Markov Decision Process for Learning Dialogue Strategies. In *Conference on Acoustics, Speech and Signal Processing.*, vol. 1 (1998), 201–204.
17. Maybury, M. T., and Wahlster, W., Eds. *Readings in intelligent user interfaces*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.
18. Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D., et al. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 23*, 3 (2015), 530–539.
19. Oviatt, S. Multimodal Interfaces. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. A. J. Sears and A., Eds. Lawrence Erlbaum Assoc., Mahwah, NJ, 2012, 413–432.
20. Pan, S. J., and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering 22*, 10 (2010), 1345–1359.
21. Pieraccini, R., and Huerta, J. Where do we go from here? Research and commercial spoken dialog systems. In *SIGdial Workshop on Discourse and Dialogue* (2005), 2–3.
22. Radomski, S. *Formal Verification of Multimodal Dialogs in Pervasive Environments*. PhD Thesis, Technische Universität Darmstadt, 2015.
23. Rudnicky, A., and Xu, W. An agenda-based dialog management architecture for spoken language systems. In *IEEE Automatic Speech Recognition and Understanding Workshop* (1999).
24. Schnelle-Walka, D., and Radomski, S. A Pattern Language for Dialog Management. In *VikingPLoP* (2012), 1–8.
25. Schnelle-Walka, D., and Radomski, S. Probabilistic Dialog Management. In *VikingPLoP* (2013), 1–12.
26. Shneiderman, B. The limits of speech recognition. *Communications of the ACM 43*, 9 (2000), 63–65.
27. Traum, D. Conversational Agency: The Trains-93 Dialogue Manager. *Workshop on Language Technology: Dialogue Management in Natural Language Systems* (1996), 1–11.
28. Turunen, M., Hakulinen, J., Räihä, K.-J., Salonen, E.-P., Kainulainen, A., and Prusi, P. Jaspis An architecture and applications for speech-based accessibility systems. *IBM Systems Journal 44*, 3 (2005), 485–504.
29. Turunen, M., Sonntag, D., Engelbrecht, K.-P., Olsson, T., Schnelle-Walka, D., and Lucero, A. Interaction and Humans in Internet of Things. In *Human-Computer Interaction (INTERACT)*, J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, and M. Winckler, Eds., Springer Berlin / Heidelberg (2015), 633–636.
30. Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D., and Young, S. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. *Empirical Methods on Natural Language Processing (EMNLP)* (2015).
31. Young, S. Using POMPDs for Dialog Management. In *Spoken Language Technology Workshop, 2006. IEEE* (2006), 8 –13.