Design and Implementation of Combined Mobile and Touchscreen-based Multimodal Web 3.0 Interfaces

Daniel Sonntag, Matthieu Deru and Simon Bergweiler
German Research Center for AI (DFKI), 66123 Saarbruecken, Saarland, Germany

Abstract - *We describe a Web 3.0 interaction system where the mobile user scenario is combined with a touchscreen-based collaborative terminal. Multiple users should be able to easily organize their information/knowledge space (which is ontology-based) and share information with others. We implemented a MP3 and video player interface for the physical iPod Touch (or iPhone) and the corresponding virtual touchscreen workbench. The Web 3.0 access allows us to organize and retrieve multimedia material from online repositories such as YouTube and LastFM.*

Keywords: User Interfaces, Mobile Interaction, Touchscreen, Design, Augmented Virtuality

1 Introduction

In addition to pure Web 2.0 interfaces, such as folksonomies and wikis, Web 3.0 interfaces bring together the social and community aspects of, e.g., user-generated content of folksonomies, with the Semantic Web [1]. One of the core intentions of the Semantic Web is providing machine understandable data for effective retrieval of commercial, scientific, and cultural data in a universal medium. Most interestingly, the idea of the Semantic Web provides new opportunities for *semantically-enabled* user interfaces. For example, in the context of Human Computing [2], anticipatory user interfaces should be human-centred and require human-like interactive functions (which includes the understanding of certain human behaviours). The understanding components shall rely on Semantic Web data structures in order to (1) transcend the traditional keyboard and mouse interaction metaphors, and (2) provide the representation structures for more complex, collaborative interaction scenarios that, most exciting, may combine mobile with terminal-based interaction and the physical with the virtual world. Over the last years, we adhered strictly to the developed rule "*No presentation without representation.*" In this paper, we discuss a combined mobile and touchscreen-based multimodal Web 3.0 interface which brings together our work in mobile scenarios, speech-based interaction, and touchscreen installations. The particular challenge we address is the implementation of a collaborative user scenario where the main application is installed on the mobile interface and a second one is installed on a touchscreen terminal. With the help of ontology-based representations, we were able to bring together those two worlds: The mobile device is recognized as a semantic item on the touchscreen (also cf. Augmented Virtuality [3]). This enables us to design a user-centred exchange terminal of multimedia data for accessing online repositories. Here, we explain our approach for implementing a combined interaction scenario. We first describe Web 3.0 interfaces in more detail (chapter 2) before introducing our Semantic Interface Element (SIE) concept (chapter 3). In our combined scenario for mobile and touchscreen-based multimodal Web 3.0 interfaces (chapter 4), we describe a Web 3.0 access which we implemented and which takes us one step closer to our targets, a complete Web 3.0 interaction. Finally, we provide a brief conclusion about our experiences in the design and implementation phase (chapter 5).

2 Web 3.0 Interfaces

Web 3.0 interfaces can either access Web 3.0 information on the Semantic Web and/or base the user interaction on Web 3.0 data.
The first point refers to an AI mashup, where a YouTube access is combined with a speech dialogue. For example, a YouTube request can be initiated by a natural speech request: „Find videos featuring Nelly Furtado". The user can retrieve the videos by natural speech and take the user ratings into account, as collaboratively constructed semantic resources. A semantic ranking system can take the ratings into account to select videos and can proactively offer extra information to the user. The second point refers to the fact that the whole interaction is based on ontologies and empirical data about the usage of the contents. This includes the ontology-based representation of the GUI layer, the discourse layer (for multimodal interaction), the data content layer (knowledge about a specific domain, e.g., the football domain, or, e.g., DBpedia), and the multimedia layer which indicated scenes on videos, regions on pictures, or headlines and summaries in texts. Integrated into a common data model (figure 1), the interaction ontology enriches the content-based fusion, selection, and suitable presentation of system reactions [4] enabled by a common semantic (ontological) framework [5].

The ontology infrastructure combines top-level ontologies about interaction models with domain-specific ontologies about the data content and the multimedia items (e.g., MPEG7) for the multimodal discourse with the user. The rendering can be done on various output devices.

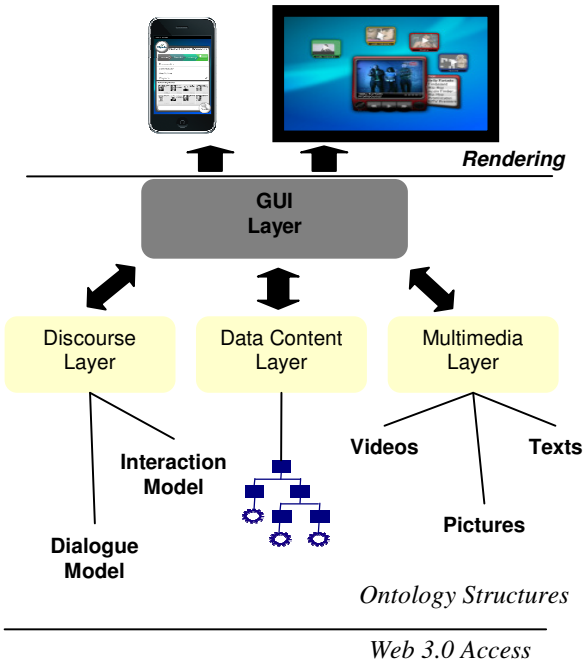

Figure 1. Common Data Model

In this work, we chose the iPod Touch (Mobile GUI) in combination with a touchscreen terminal (GUI, Haptics, and Speech). Physical elements, e.g., physical objects or people, are dynamically integrated into the virtual world: physical devices dynamically exchange information with the touchscreen based GUI. Each physical interaction is interpreted as a gesture. Thanks to various surface forms for rendering (e.g., Liquid List or Faceted Lists), every physical interaction device gets a corresponding clone on the touchscreen. With the help of this clone in the virtual world, we have the chance to interact with the virtual world in real-time using advanced interaction possibilities (Touchscreen and Speech).

3 Semantic Interface Elements (SIEs)

We implemented an abstract container concept called *Semantic Interface Elements* (SIEs) for the representation and activation of multimedia elements visible on the touchscreen. (Technically, they also trigger requests to Web Services.) Semantic-based data integration frameworks for multimedia data allows authors to write interactive multimedia presentations, and are very valuable in planning environments where the complete content displayed on screen is dynamically build up. A further objective is a description framework for the semantic output processing, which can be used on different end-user devices. The presentation ontology specifies interaction patterns relevant for the application. Currently the graphical user interface driven by the SIE concept is enabling the following semantic presentation elements based on ontology descriptions (OWL):

- Tables with selectable cells
- Lists with selectable elements and Faceted Browsing
- Graphs with nodes, which can be selected
- Containers for elements of lists, tables, graphs, and several different media types

The containers are further abstracted for the representation and activation of multimedia elements visible on the touchscreen. We call the implementations of the SIEs for our music environment *Spotlets*. To implement a SIE hierarchy, a general interaction event hierarchy (e.g., for "select", "drag & drop", or "drag-to-function" commands) is needed.[1] Since SIEs are intelligent agents as semantic objects that perform different actions controlled by gestures or speech, they enable several functions to be called. These functions might be called if a user interacts with the SIE implementation in a certain way. One of the domain SIEs (a spotlet for the video clips), for instance, tries to find artists similar to that from an MP3 that is dragged on top of it.

Each SIE has an interaction zone or area; the interaction zone defines the action behaviour and serves as *input* or *semantic drop-zone*. (Currently, the spotlets are using object drop as an input.) Also in the case of an MP3 (we consider this object/file as semantically annotated), a search can be made by dropping the object onto a spotlet in order to retrieve similar songs or artists.

After a SIE's action is used, an output signal is provided. In this way, a mashup of SIEs can be created. We distinguish *terminal* and *mobile* SIEs. Terminal SIEs interact directly with the terminal (processing layer), making the display, browsing, and handling of large amounts of data a definite possibility. Mobile SIEs serve for user personalization and for the handling of text, audio, and video data on the mobile client. We do not display more than one mobile SIE concurrently. The SIEs use an XML-based specification language for the instantiation of a SIE and the description of its contents. Within the dialogue-system, this XML is parsed to build up ontology instances of the presentation items and the displayed content of the SIEs.
The graphical interface of the touchscreen client and the speech interface for the touchscreen, make up, together with the mobile devices, the combined mobile interaction scenario to organize a knowledge/information space. Whereas other architectures for mobile Web Service content have already been proposed (e.g., for business purposes [6]), we focus on user-generated multimedia content and the mobile/terminal exchange environment as a unique contribution.

[1] For the GUI surface was implemented as Flash content to be locally executed on the touchscreen device. We used Flash 10 with Actionscript 3 which is one of the industry standard tools for creating visual content and applications.

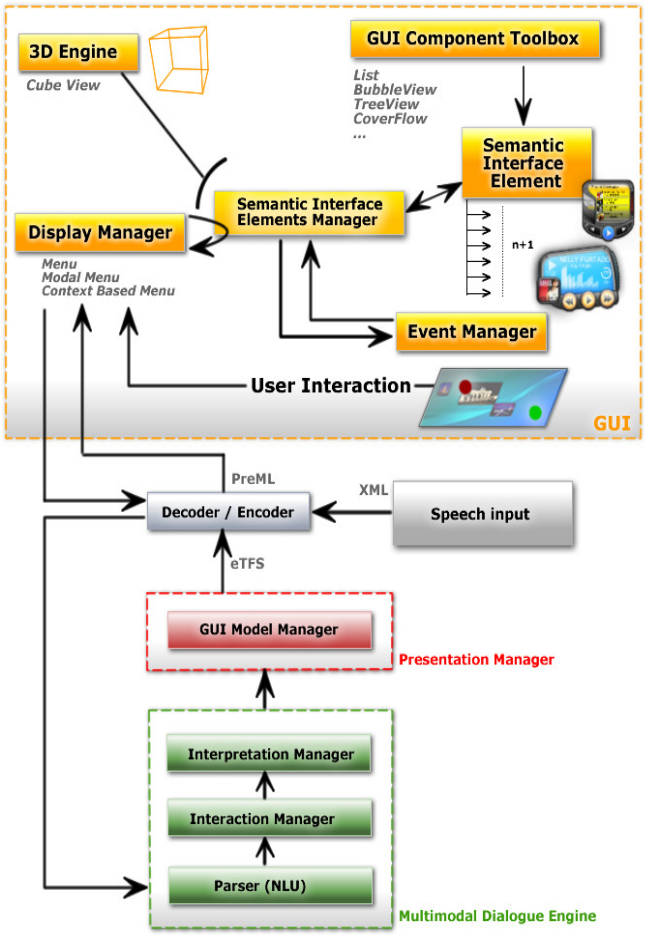

Figure 2. Technical SIEs Architecture

The technical SIE architecture comprises of several GUI-related modules (such as the Semantic Interface Elements Manager or the Event Manager), dialogue-engine-related modules (such as the Interaction Manager or natural language Parser), and the Presentation Manager sub-module (i.e., the GUI Model Manager). The most important part of the architecture is the Display Manager which observes the behaviour of the currently displayed SIE (e.g., whether it was moved by the user or dropped somewhere). The display manager dispatches XML based messages to the dialogue system with the help of a message Decoder / Encoder. The Multimodal Dialogue Engine then processes the requests in the Interaction and Interpretation Manager modules. A new system action or reaction is then initiated and sent to the Presentation Manager. The GUI Model Manager builds up a presentation message in eTFS notation (internal typed feature structure format), before the complete message is sent to the Display Manager as PreML message (a presentation markup language in a XML dialect for presentation messages initiated by the dialogue system). The Presentation Manager also stores the display model while using the model-view-controller design pattern (the Display Manager is the view and the Multimodal Dialogue Engine the controller).

4 Combined Interaction Scenario

In the context of a large-scale research project (THESEUS) we implemented a prototype of our multimodal WEB 3.0 interface as Collaborative Media Exchange Terminal. This terminal is designed to be installed in a public place (or a café) and especially emphases the exchange of free music samples or video trailers which you can find on the Web. This is beneficial because isolated mobile interaction devices have the following drawbacks:

- Often, direct software downloads are slow and cost-intensive.
- On mobile devices, searching for information in large information spaces, such as online music and video repositories, is tedious. This is due to the fact that the interaction possibilities (e.g., keyboard) are quite limited, the screen is too small to display complex relationships between media items, touchscreen and haptic interactions are often limited to selection commands, and last-but-not-least, natural speech-based interaction is not possible without a server-based speech dialogue platform.
- People cannot easily share and exchange their files directly from one mobile device to the other (due to incompatible wireless technologies, e.g., WLAN and Bluetooth). The design of direct exchange programs for mobile devices leaves a lot to be desired.

In order to address these issues, we envisioned to install the Collaborative Media Exchange Terminal in a in a public place (or a café). Internet access can easily be provided. A special microphone allows us to isolate the speech command when the users are standing in front of the touchscreen table. And, intuitive user-centred interaction should allow every user to easily exchange music files. In this context, user-centered design is to be understood as the need to allow a first-time user to understand the touchscreen interaction principles, and that each user's learning curve is steep.

In addition, we paid very much attention to the speech input (first and foremost for posing the natural language query or command) since the terminal works without a keyboard and speech input is most natural in a collaborative user setting. Of course, any user-centred design benefits from natural speech input if the speech recognition and understanding rate is high. (We also experimented with a multi-device solution for situations where the ASR component is rather limited. For example, person names for asking questions to Wikipedia and DBpedia are hard to understand because of many open vocabulary words.) In the following, we explain the touchscreen GUI and haptics interactions, and speech-based interactions in more detail (additionally, a YouTube video is available: http://www.youtube.com/watch?v=hAAwKxeoCrk).

4.1 Touchscreen GUI and Haptics

The touchscreen scenario combines a multi-touch screen with several graphical GUI widgets, the SIEs (figure 2), as intelligent agents. As can be seen, multiple music and video SIEs (here called spotlets) can be arranged on the surface, whereas only one of them is in visual (and discourse) focus.

In addition to the YouTube connection for providing videos and user ratings where we display the best-5 answer videos according to the String-based standard search API[2], we were also able to connect to the Seeqpod REST API[3], and the equivalent APIs for Lastfm[4] and Flickr[5]. This means that multimedia items from all those sources can be arranged, displayed, and manipulated on the touchscreen.

Figure 3. YouTube Spotlets

In addition, the multimedia contents from YouTube and Flickr can be dragged onto the mobile device (cf. figure 5).

4.2 Speech-based Interactions

Speech dialogue systems are very different from more traditional interactive systems whose usability aspects have been investigated for decades, such as systems controlled through graphical user interfaces involving screen, keyboard, and mouse.

In previous work [7], we demonstrated their usage on mobile clients for multimodal interaction with ontological knowledge bases and semantic Web Services. In the context of Web 3.0 interaction, we decided to implement speech interaction only for the collaborative touchscreen terminal.

Speech is perceptually transient rather than static. However, its usage is very powerful for semantic search scenarios because the speech recognition and understanding components can easily update their grammars according to a new semantic service. For example, each semantic Web 3.0 service has a proper name tag associated with it. When a user says: "Address Flickr", we already have updated the grammar which understands Flickr only as a service name.

The dialogue platform (available in the commercial product ODP at SemVox, see http://www.semvox.de/) is used for the interpretation of multiple automatic speech recognition (ASR) hypotheses. These are then stored, just as the user clicks, in ontological representation of the user's input. According to this representation, an appropriate system reaction is planned on a modality-independent level. A multimodal fission component is then used to select the contents that are going to be presented on the output channels, in a modality-dependent way (natural speech generation or the instantiation of a SIE instance on the touchscreen GUI). Finally, a multimodal generation component generates an appropriate surface representation. The GUI-related presentation part is sent to the presentation manager as illustrated in figure 2. Figure 3 shows the touchscreen with the multimedia contents and a MP3 player spotlet.

Figure 4. Touchscreen Surface and MP3 Player Spotlet

4.3 Touchscreen Installation

Concerning multimodal result presentations [8], we developed specific guidelines according to the user-centred design and usability criteria: In our envisioned touchscreen interaction scenario where we address the Web 3.0 data layer,

(1) Many results are not naturally presented auditorily, which means that we try to exploit the graphical display as much as possible. Accordingly, all results will be presented on the screen as SIE contents;

[2] http://code.google.com/apis/youtube/overview.html
[3] http://www.seeqpod.com/api.php
[4] http://www.lastfm.de/api/intro
[5] http://www.flickr.com/services/api/

(2) Every multimodal dialogue act, such as presenting answers or giving immediate feedback on a user command or query, is realized as a graphical element (for example, we implemented a micro/speaker spotlet to indicate the status of the ASR/Speech Synthesis);

(3) We hold that longer speech syntheses (even in high quality) are distracting. Generally, speech synthesis belongs to the more obtrusive output channels. On this note, we propose an integration of graphics and speech output to keep acoustic messages short and simple. (Often, the media results are played anyway.)

Data exchange is accomplished between terminal SIEs and mobile SIEs. A gesture on the screen can serve to hand over the data. A user can place his iPod on the surface of the terminal touchscreen. The displayed media items (e.g., MP3 files) are all ID3[6]-tagged and supply meta information about the artist, the genre, and the album. These metadata is used to build up semantically annotated objects.

These objects are displayed in form of, e.g., a *liquid list* (figure 4). A video camera captures the position of the mobile interfaces it recognizes as SIE (figure 5).

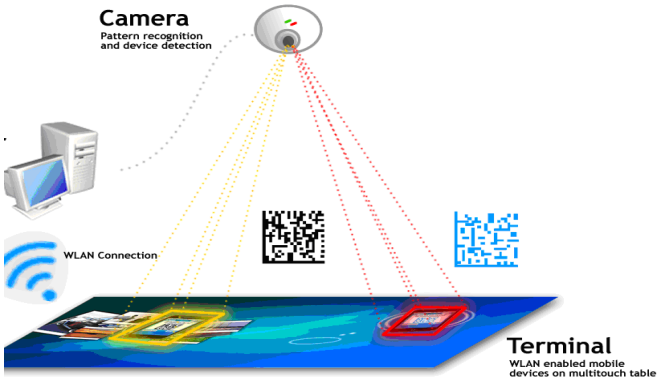

Figure 5. Combined Scenario and Recognition of Devices

When the mobile device is detected by the camera[7], a circle appears around it and a category list of the data extracted from the mobile device is arranged in the interaction zone around it. Each generated icon is a symbol for the concrete files stored in a shared folder on the mobile device.

Each of these items (semantic objects) can then be associated with a terminal SIE or exchanged between several iPods or iPhones by using "drag-&-drop" (figure 6).

[6] ID3 is a metadata container and allows us to store information such as the title, artist, album and track number in the media file itself.

[7] We use a two-dimensional barcode on the screen of the iPod/iPhone. See http://www.barcodesoft.com/datamatrix_font.aspx for more information about the Data Matrix Barcode software.

Figure 5. Collaborative SIEs Scenario

In the resulting, combined scenario, users (U:) have stored their personal multimedia information, which they want to share with each other on their iPods or iPhones. In addition, the users wish to download additional video content. These (collaborative) tasks can be fulfilled while using the touchscreen terminal system (S:) and natural dialogue, as exemplified in the following combined multimodal dialogue which we implemented:

1. U: Start client applications and register several iPods on the terminal. (WLAN IP connection)

2. U: Place, e.g., 2 registered iPods on the touchscreen (ID3 tagged)

3. S: Shows media circles for iPod contents

4. U: Can select and play videos and search YouTube and Lastfm etc. by saying:

   "Search for more videos and images of this artist."

5. S: Delivers additional pictures, video clips and replies:

   "I found 40 videos and selected 5. Do you want me to play them?"

6. U: ***"Yes, play the first two hits."***

7. S: Initiates two video-spotlets and plays the videos.

8. U: Drag image of *Nelly Furtado* onto second iPod; Remove second iPod (figure 5, right) from the touchscreen.

9. S: The media is transferred via IP connection.

5 Conclusions

Multimodal speech-based interfaces offer great opportunities when designing mobile human computer interfaces. We implemented a collaborative user scenario where one application is installed on the mobile interface and a second one is installed on a touchscreen exchange terminal. By rigorously following our presentation and representation principle "*No presentation without representation*", we can refer to all presentation elements at input processing as a side-effect of profound data structure design. Multitouch, drag-&-drop, and speech input can be used. This makes the interaction between the virtual world on the touchscreen and the physical iPod or iPhone a little more seamless.

In future work, we will address more robustness and flexibility in speech recognition and understanding for more complex dialogue scenarios, more SIEs (e.g., we are working with DBpedia contents), and the exploration of more fine-grained co-ordinated and synchronized multimodal presentations in mobile environments. The DBpedia spotlet should allow the user to browse information on the big touchscreen by following semantic relations, before he or she selects some interesting articles about the explored contents (Wikipedia contents) which can then be uploaded on the iPod or iPhone. Equipped with the DBpedia/Wikipedia contents, our touchscreen installation would also be suitable for museum entrance halls or other public places where people should be able to look up facts and store information about artists on their mobile devices. A longer-term research question we ask is whether speech commands on the mobile client interfaces would also be a benefit in the context of the exchange terminal application, as demonstrated in the context of, e.g., navigation maps [9] on mobile devices while on the go.

6 Acknowledgments

This research has been supported by the THESEUS Programme funded by the German Federal Ministry of Economics and Technology (01MQ07016). The responsibility for this publication lies with the authors.

7 References

[1] Fensel, D., Hendler, J.A., Lieberman, H., Wahlster, W.: Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, Cambridge (2005)

[2] Huang, T. S., Nijholt, A., Pantic, M., & Pentland, A. (eds.): Artifical Intelligence for Human Computing Artifical Intelligence for Human Computing, Springer, LNAI 4451, (2007)

[3] Milgram, P., Takemura, H., Utsumi, A., Kishino, F.: Augmented Reality: A class of displays on the reality-virtuality continuum. Proceedings of Telemanipulator and Telepresence Technologies (1994)

[4] Wahlster, W.: Planning multimodal discourse. In: Proceedings of the 31st annual meeting of ACL, Morristown, NJ, USA, pp. 95–96, (1993)

[5] Geurts J., Bocconi S., van Ossenbruggen J., & Hardman L.: Towards ontology-driven discourse: From semantic graphs to multimedia presentations. In: Second International Semantic Web Conference, Sanibel Island, FL, USA (2003)

[6] Chen, M., Zhang, D., & Zhou, L.: Providing web services to mobile users: the architecture design of an m-service portal. International Journal of Mobile Communications, 3(1), 1-18, (2005)

[7] Sonntag, D., Engel, R., Herzog, G., Pfalzgraf, A., Pfleger, N., Romanelli, M., Reithinger, N.: SmartWeb Handheld - Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services (extended version). Thomas Huang, Anton Nijholt, Maja Pantic, Alex Plentland (eds.): LNAI Special Volume on Human Computing, Volume 4451, Springer (2007)

[8] S. Oviatt.: Ten myths of multimodal interaction. Communications of the ACM, 42(11):74–81 (1999)

[9] Sonntag, D.: Context-Sensitive Multimodal Mobile Interfaces, Speech and Gesture Based Information Seeking Interaction with Navigation Maps on Mobile Devices. SIMPE Workshop, In: Proceedings of MobileHCI, (2007)