A Multimodal Dialogue System for Medical Decision Support in Virtual Reality

Alexander Prange, Margarita Chikobava, Peter Poller, Michael Barz, Daniel Sonntag
German Research Center for Artificial Intelligence (DFKI)
66123 Saarbrücken, Germany
{Firstname}.{Lastname}@dfki.de

Abstract

We present a multimodal dialogue system that allows doctors to interact with a medical decision support system in virtual reality (VR). We integrate an interactive visualization of patient records and radiology image data, as well as therapy predictions. Therapy predictions are computed in real-time using a deep learning model.

1 Introduction

Modern hospitals and clinics rely on digital patient data. Simply storing and retrieving patient records is not enough; in order for computer systems to provide interactive decision support, one must represent the semantics in a machine readable form using medical ontologies (Sonntag et al., 2009b). In this paper, we present a novel real-time decision support dialogue for the medical domain, where the physician can visualize and interact with patient data in an virtual reality environment by using natural speech and hand gestures.

Our multimodal dialogue system is an extension of previous work by Luxenburger et al. (Luxenburger et al., 2016) where we used an Oculus Rift with an integrated eye-tracker in a medical remote collaboration setting. First, the radiologist fills out a findings form using a mobile tablet with a stylus. The data is then transcribed in real-time using automated handwriting recognition, parsed, and represented based on medical ontologies. Then, the doctor, or any other health professional, enters virtual reality and interacts with patient records using the multimodal dialogue system. Through the temporal synchronization of visual and auditory events in VR, we support multisensory integration (Morein-Zamir et al., 2003). This way we profit from superadditivity (Oviatt, 2013) to further enhance multisensory perception.

2 Architecture

Modern hospitals and clinics are highly digitalized; in order to integrate our system seamlessly into everyday processes, we designed a highly flexible architecture, which can be connected to existing hospital systems (e.g., PACS, a picture archiving and communication system) and connects novel interaction devices such as VR glasses and head-mounted displays (HMDs). As depicted in Figure 1, all devices in this scenario are either connected directly or through adapters to the *Proxy Server* using XML-RPC, a remote procedure call protocol which uses XML to encode information that is sent via HTTP between clients and server. The Proxy Server manages and relays the cross-platform communication between the different devices. The mobile device for instance retrieves patient data and medical images through the Proxy Server from the hospital's *PACS* and *RIS* (radiology information system). The doctor then fills out the report, and the results are send back. Some components, like the PACS and RIS, are not connected directly through XML-RPC to the rest of the system, but through the *Patient Data Provider*, which provides an abstraction layer to the other devices. This way we can ensure a flexible integration of different proprietary software solutions that are already being used in hospitals.

2.1 Mobile Device

Even though modern hospitals and clinics are highly digitalized, there are many everyday processes that are still performed using pen and paper. Our approach in this scenario is based on the work of Sonntag et al. (Sonntag et al., 2014) where they use digital pens to improve reporting practices in the radiology domain. Instead of using a digital pen on normal paper, we create a fully digital version of the radiology findings form (in

Proceedings of the SIGDIAL 2017 Conference, pages 23–26,
Saarbrücken, Germany, 15-17 August 2017.

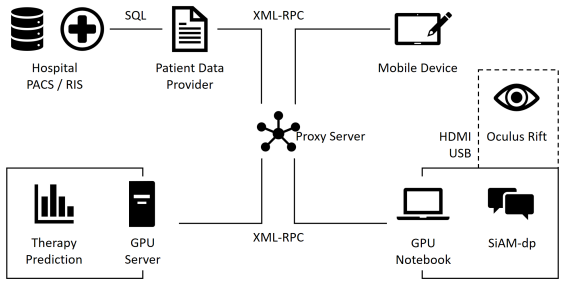

Figure 1: Architecture diagram

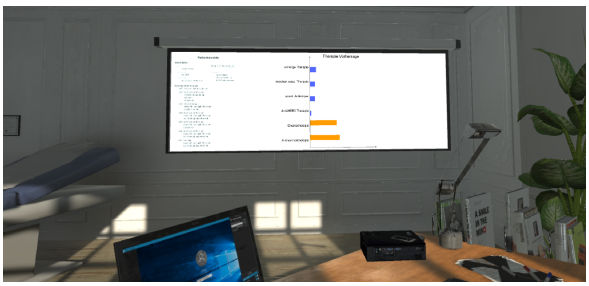

Figure 2: Screenshot of therapy prediction results inside virtual reality

this case mammography), to be used on a mobile device with integrated stylus. The radiologist writes the report directly onto the tablet using the stylus and through real-time handwriting, gesture, and sketch recognition, the entire content is transcribed, exported and written into the hospital's database. Our approach has several advantages over the traditional form filling process: (1) the contents are instantly transcribed and parsed into concepts of medical ontologies, (2) real-time feedback about the handwriting recognition process allows for a direct validation of input data, and (3) medical images are taken directly from the hospital's PACS, are then displayed on the screen and can be annotated by the radiologist. We use the Samsung Note series as mobile devices, because they feature a special Wacom digitiser technology for the stylus input; we built our software on top of the MyScript[1] handwriting recognition engine.

2.2 Virtual Reality

We created a Unity3D application[2] that resembles a real world doctor's office. The user can move freely inside the room using positional tracking and may also look around using head tracking. To enable immersive and remote interaction with medical multimedia data, we use a projection on the wall, where the patient files, the previously annotated digital form, and the therapy predictions are shown (see Figure 2). Navigation inside the documents, like zooming or scrolling through pages, can be achieved either through natural speech interaction or by using the Oculus Touch controllers, which we render as hands inside VR.

[1] http://myscript.com/
[2] http://unity3d.com/

2.3 Decision Support

Our medical dialogue system facilitates support for deciding which therapy is most suitable for a given patient. We integrated a prediction model for clinical decision support based on deep learning (Esteban et al., 2016) as backend service running on a dedicated GPU server. They presented a recurrent neural network (RNN) to include dynamic sequences of examinations which was modified to take dynamic patient data as additional input. This model was trained on a set of structured data from 475 patients, containing a total of 19438 diagnoses, 15352 procedures, 59202 laboratory results and 13190 medications. All personal data, such as names, date of birth, patient-IDs were anonymized accordingly and all date and time references were shifted. For our dialogue system, fast response times are of particular interest. We use TensorFlow (Abadi et al., 2016) to enable GPU-accelerated predictions on a scalable platform. Our service runs on a dedicated high-performance computer and is accessible to the dialogue system through our Proxy Server.

3 Dialogue

In order to facilitate coordinated interactions on the patient data within the virtual reality environment, we developed a multimodal dialogue interface that allows us to operate and interact by speech and gestures. The multimodal dialogue system supports three different types of interactions: (1) interactions with the patient data shown on the virtual display (e.g., *"Open the patient file for Gerda Meier.", "Show the next page."*); (2) interactive question answering (QA) about the contents of a patient record (e.g., *"When was the last examination?"*); and (3) control of the therapy prediction component (e.g., *"Which therapy*

is recommended?"). Within the dialogue the following speech interactions and phenomena are realized:

- Navigation inside patient records (e.g., open/close file, scroll, zoom, turn page)
- Anaphoric reference resolution (e.g., *"What is her current medication?"*)
- Elliptic speech input (e.g., *"... and the age?"*)
- Multimodal (deictic) dialog interactions (e.g., *"Zoom in here"* + [user points on a region on the display])
- Cross-modal reference resolution (e.g., *"What is the second best therapy recommendation?"*)

3.1 Dialogue Implementation

The implementation of the dialogue follows the rapid engineering principles (Sonntag et al., 2009a) and is implemented with SiAM-dp (Neßelrath, 2015), an open development platform for multimodal dialogue systems. All knowledge representations and dialogue structures follow a declarative specification with ontology structures. First, the already existing patient data model of the patient database was mapped onto the corresponding domain ontology for SiAM-dp's knowledge manager, which is initialized with the specific patient instances at the beginning of each dialogue session. The speech recognition grammar is loaded into Nuance's speech recognizer[3].

The dialogue model is based on finite-state machines; the mapping of user intentions to matching multimodal system reactions is defined declaratively. The determination of the user intention in SiAM-dp follows a fusion process: SiAM-dp's modality specific user input analysis components (speech recognition, gesture analysis) and their fusion in conjunction with reference resolution within the discourse manager. The realization of multimodal output (speech output, virtual display content modifications, therapy prediction invocation) is coordinated by SiAM-dp's presentation planning component. The software itself runs on the same machine that has the Oculus Rift and the Touch controllers attached. Technically speaking, SiAM-dp is operated with standard speech recognition and synthesis (Nuance, SVOX), connected to Oculus Rift's microphone and speakers as audio input and output devices. An example dialogue is as follows:

U.1 *"Show the patient file for Gerda Meier."*
S.1 *"Here is the patient file for Gerda Meier."* [patient data is displayed on the display inside the VR room]
U.2 *"What was the last examination?"*
S.2 *"Mrs. Meier recently received a mammography."*
U.3 *"When was it?"*
S.3 *"The mammography was made on the 10th of March."*
U.4 *"Now show me the patient file for Paula Fischer."*
S.4 *"Here is the patient file for Paula Fischer."* [new patient data is displayed]
U.5 *"Zoom in here."* [user points on a region on the display using the Oculus Touch controller]
S.5 [virtual display is zoomed accordingly]
U.6 *"Which therapy is recommended?"*
S.6 *"For Paula Fischer chemotherapy is recommended."* [bar chart with therapy prediction is displayed]

In (U.1) the user requests a patient file to be presented on the display inside the VR room. The corresponding system output (S.1) is multimodal: speech output is synchronized with the presentation of the patient file. The user then requests information about the patient data currently shown on the display (U.2), e.g. anamneses and previous therapies. This user input contains an ellipsis: the name of the patient is not mentioned. SiAM-dp's discourse manager resolves it from the dialogue context that was filled in (U.1). Further questions about specifics may be asked (U.3). The context infers that "it" refers to the mammography just mentioned (rule-based anaphora resolution).

The next utterance (U.4) shows that users may shift the topic at any point, for instance by requesting other patient data. (U.5) is an example of a multimodal input consisting of a speech input and a corresponding pointing gesture. Processing this user input is only possible if both modalities are in a certain time frame and correctly fused.

The main dialogue move is (U.6), as it triggers the real-time therapy prediction process on the GPU Server. The system's response in (S.6) is again multimodal as the requested therapy is presented on the virtual display, together with synthesized speech output.

[3] http://www.nuance.com

Anaphora resolution is also handled in our system. Since the patient file represented on the display is always synchronized with the current discourse model and within SiAM-dp depending on the context modelled as discourse memory (Sonntag, 2010) the system can resolve utterances like *"when was her last examination?"*

4 Conclusions and Future Work

In this paper, we presented our multimodal dialogue system implementation in virtual reality. It provides a first example of an automated decision support system that computes therapy predictions in real-time using deep learning techniques. Our multimodal dialogue system, in combination with interactive data visualization in virtual reality, is meant to provide an intuitive dialogue component for helping the doctor in his or her therapy decision. Preliminary evaluations in the clinical data intelligence project (Sonntag et al., 2016) are encouraging and we believe that such multimodal-multisensor interfaces in VR can already be designed and implemented to effectively advance human performance in medical decision support.

Currently we are investigating how displaying complex 3D medical images (e.g., DICOM) in VR can improve the diagnostic process. We are also looking into possibilities to include additional input modalities such as gaze information from eye-tracking to further improve the multimodal interaction.

As an extension to the dialogue it is planned to include ambiguity resolution by asking clarification questions. If a patient name is ambiguous, the system could ask for clarification (U: "Open the patient file of Mrs. Meier." S: "Gerda Mayer or Anna Maier?"). In addition users should be able to change or add patient data through natural speech.

Acknowledgments

This research is part of the project "clinical data intelligence" (KDI) which is founded by the Federal Ministry for Economic Affairs and Energy (BMWi).

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. USENIX Association, GA, pages 265–283.

C. Esteban, O. Staeck, S. Baier, Y. Yang, and V. Tresp. 2016. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. pages 93–101.

Andreas Luxenburger, Alexander Prange, Mohammad Mehdi Moniri, and Daniel Sonntag. 2016. Medicalvr: Towards medical remote collaboration using virtual reality. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, New York, NY, USA, UbiComp '16, pages 321–324.

S. Morein-Zamir, S. Soto-Faraco, and A. Kingstone. 2003. Auditory capture of vision: examining temporal ventriloquism. *Brain Res Cogn Brain Res* 17(1):154–163.

Robert Neßelrath. 2015. *SiAM-dp : An open development platform for massively multimodal dialogue systems in cyber-physical environments*. Ph.D. thesis, Universität des Saarlandes, Postfach 151141, 66041 Saarbrücken.

S. Oviatt. 2013. *The Design of Future Educational Interfaces*. Taylor & Francis.

Daniel Sonntag. 2010. *Ontologies and Adaptivity in Dialogue for Question Answering*, volume 4 of *Studies on the Semantic Web*. IOS Press.

Daniel Sonntag, Gerhard Sonnenberg, Robert Nesselrath, and Gerd Herzog. 2009a. Supporting a rapid dialogue engineering process. In *Proceedings of the First International Workshop On Spoken Dialogue Systems Technology. International Workshop On Spoken Dialogue Systems Technology (IWSDS-2009), December 9-11, Kloster Irsee, Germany*. o.A.

Daniel Sonntag, Volker Tresp, Sonja Zillner, Alexander Cavallaro, Matthias Hammon, André Reis, Peter A. Fasching, Martin Sedlmayr, Thomas Ganslandt, Hans-Ulrich Prokosch, Klemens Budde, Danilo Schmidt, Carl Hinrichs, Thomas Wittenberg, Philipp Daumke, and Patricia G. Oppelt. 2016. The clinical data intelligence project. *Informatik-Spektrum* 39(4):290–300.

Daniel Sonntag, Markus Weber, Alexander Cavallaro, and Matthias Hammon. 2014. Integrating digital pens in breast imaging for instant knowledge acquisition. *AI Magazine* 35(1):26–37.

Daniel Sonntag, Pinar Wennerberg, Paul Buitelaar, and Sonja Zillner. 2009b. Pillars of ontology treatment in the medical domain. *J. Cases on Inf. Techn.* 11(4):47–73.