TOWARDS INCORPORATING 3D SPACE-AWARENESS INTO AN AUGMENTED REALITY SIGN LANGUAGE INTERPRETER

Fabrizio Nunnari, Eleftherios Avramidis, Vemburaj Yadav, Alain Pagani, Yasser Hamidullah, Sepideh Mollanorozy, Cristina España-Bonet, Emil Woop, Patrick Gebhard

German Research Center for Artificial Intelligence (DFKI)

ABSTRACT

This paper describes the concept and the software architecture of a fully integrated system supporting a dialog between a deaf person and a hearing person through a virtual sign language interpreter (aka avatar) projected in the real space by an Augmented Reality device. In addition, a Visual Simultaneous Localization and Mapping system provides information about the 3D location of the objects recognized in the surrounding environment, allowing the avatar to orient, look and point towards the real location of discourse entities during the translation. The goal being to provide a modular architecture to test single software components in a fully integrated framework and move virtual sign language interpreters beyond the standard "front-facing" interaction paradigm.

Index Terms— sign language, machine translation, augmented reality.

1. INTRODUCTION

Immersive technologies such as Augmented Reality (AR) have attracted a lot of interest recently, with several applications being already commercially available. Although there has been a multitude of development directed to the needs of hearing users, little has been developed for deaf and hard-of-hearing people, whose main language of communication is the sign language (SL). Since AR has video as one of its main modi, 3D virtual agent (a.k.a. avatar) capable of signing can be displayed in the user's eyesight and convey information to the user.

Additionally, the ability of AR to be "placed" in and interact with the real world makes it eligible for assistant applications, where the deaf or hard of hearing user can can interact with the agent with regards to the current situation or the real space they are currently in. Relevant applications could be provision of directions in the real space or virtual spoken-to-sign language interpretation, where the virtual interpreter appears next to the original hearing speaker [1].

The research reported in this paper was supported by BMBF (German Federal Ministry of Education and Research) via the project SocialWear (Socially Interactive Smart Fashion, grant no. 01IW20002) and by the European Union via the project European Language Equality 2 (grant no. LC-01884166 – 101075356 ELE2).

In this paper we aim to describe the development plan of an AR prototype that could be adapted to the needs of sign language. Since the grammar of sign languages employs several spatial features, we argue that the awareness of the real space around the user should be used to enable and feed this features. This way, under the umbrella of the technological platform of AR, one would have to combine several state-of-the-art technologies, such as Augmented Vision and Sign Language Generation via Natural Language Processing and Computer Graphics.

The goal of the development of this prototype is threefold. First, implement a test bed to verify that the single components are mature enough to be employed in a real-case scenario, beyond the comfort of in-lab metric-driven tests. Second, verify with the community (through feedback collection) the effectiveness of the use of AR devices for the projection of SL interpreters. Third, through a set of user studies, verify the hypothesis that space-aware virtual interpreters (able to direct deictict gestures, eye gaze, and body orientation towards points of interest) are better understandable and desired than existing approaches where signing avatars are limited to frontal communication towards the framing camera.

2. RELATED WORK

There exist only a few research prototypes for sign language synthesis (e.g., JASigning [2], EMBR [3], and Paula [4]), all synthesising the avatar on a screen window, where the avatar is forward-facing the virtual camera.

In addition to screen-rendered avatars, a number of more recent experiments started to propose the projection in virtual or augmented reality (AR). Using AR departed from experiments where sign language users favored a hands-free device that allowed having simultaneous eye contact with the speaker while being able to use their hands, as shown by [5] who compared the usability of a head-mounted device with a smartphone while being connected to a live interpreter. [6] developed a holographic avatar that translates real-time English to Signed English on AR glasses, and was tested on deaf children during math classes at school. [1] presented a proof-of-concept evaluation for a system that provides translation of speech to virtually performed sign language on AR glasses,

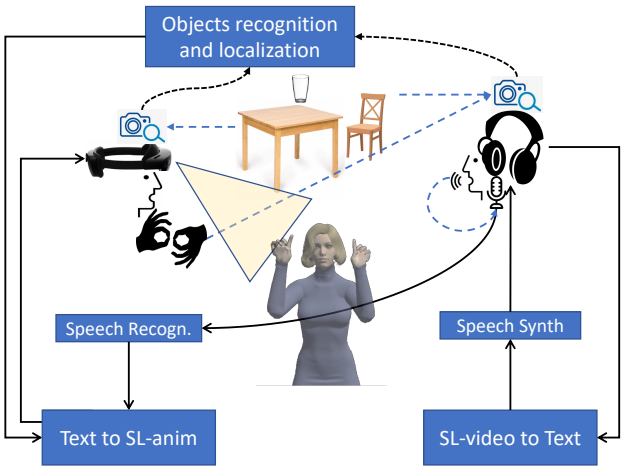

Fig. 1. An abstract view of the prototype.

which received a high acceptance rate among deaf and hard-of-hearing people. Several works use AR to aid SL learning [7, 8], e.g., by using a Kinect [9].

However, most work in signing avatars and/or AR is in preliminary stage and a considerable amount of research is required to bring them to a state that is usable and fully accepted by the sign language community. None of them considered the idea we are suggesting in this paper: to take advantage of the context information from the surrounding environment and using it as shared space (between the avatar and the deaf user) during the synthesis of the SL utterances.

3. CONCEPT

Figure 1 shows a top-level abstract view of the concept. We envision that a deaf user can wear an AR device able to project a virtual interpreter on the top of the real environment. The same device is equipped with a camera sensor capturing information about the surrounding environment. In the same environment, the speaking partner wears a headset augmented with a camera also scanning the surrounding environment. Whenever the speaking partner wants to communicate, his voice is translated first into text and then into the animation data needed to animate the virtual interpreter. On the other direction, when the deaf users signs something, the camera of the augmented headphones records the video, which is translated into text, and finally synthesised as voice and played back in the headphones.

Both cameras mounted on the head of the two users continuously take snapshots of the surrounding environment. The images are sent to a software module able to perform object recognition and classification of the surrounding objects, together with the 3D reconstruction the surrounding 3D space and hence associate each recognized object with a 3D position in space. The space localization data will be continuously updated for the Text-to-SL-animation translation module, which will animate the virtual interpreter in a "space-aware" fashion. The signing avatar will be then able to inflect the signs according to their real position in space. This will potentially affect deitctic gestures (e.g., INDEX), verbs expressing motion and transfer (e.g., to walk, to give), and eye gaze.

4. SOFTWARE DESIGN

Figure 2 depicts the software architecture proposed to realize a working prototype, which will integrate a set of state-of-the-art software modules specialized in their own sub-tasks.

4.1. Interaction design

As automatic segmentation of sign language or spoken sentence is not 100% reliable, and out of the scope of this work, we decided to give full control to the interaction to the user through the use of explicit gestures or devices.

When the deaf user wants to communicate, he performs an activation gesture in front of his/her AR device. This will give an acoustic signals to the hearing users which can then turn his camera towards and signing user and start recording the motion. Similarly, when the speaking users wants to communicate, he will press a button on a portable device. This will give a visual feedback on the AR device of the deaf user, which can know in advance the communication intent and observe the speaking user, to grasp more communicative information through additional communication channels (e.g., lip reading, body posture, facial expressions) even before the activation of the signing avatar.

4.2. Speech Synthesis and Recognition

We do not plan to put direct effort in the development of speech synthesis (text-to-speech) or recognition (speech-to-text) systems. A number of open source alternatives are available that can be wrapped for a networked integration. As for speech synthesis, we will consider if relying on long-lasting robust and lightweight engines based on a statistical approach, such as MaryTTS [10], or on more modern neural based approaches, as in Tacotron [11]. Concerning speech recognition, possible candidates are Vosk [12], and Whisper [13].

4.3. Avatar Manager

When translating from speech to sign language, the goal of this module is to perform the following sequence of operations: get the signal from the headset that there is a voice sentence to translate, ask the speech recognition module to convert the sound into text, ask the Text Conversor module to convert the text into MMS (multi-modal signstream), gather information about the spatial location of objects, ask the Animation Synthetizer to convert the space-augmented gloss sequence into an animation, and finally send the animation data to the Avatar Player (running on the deaf user AR device).

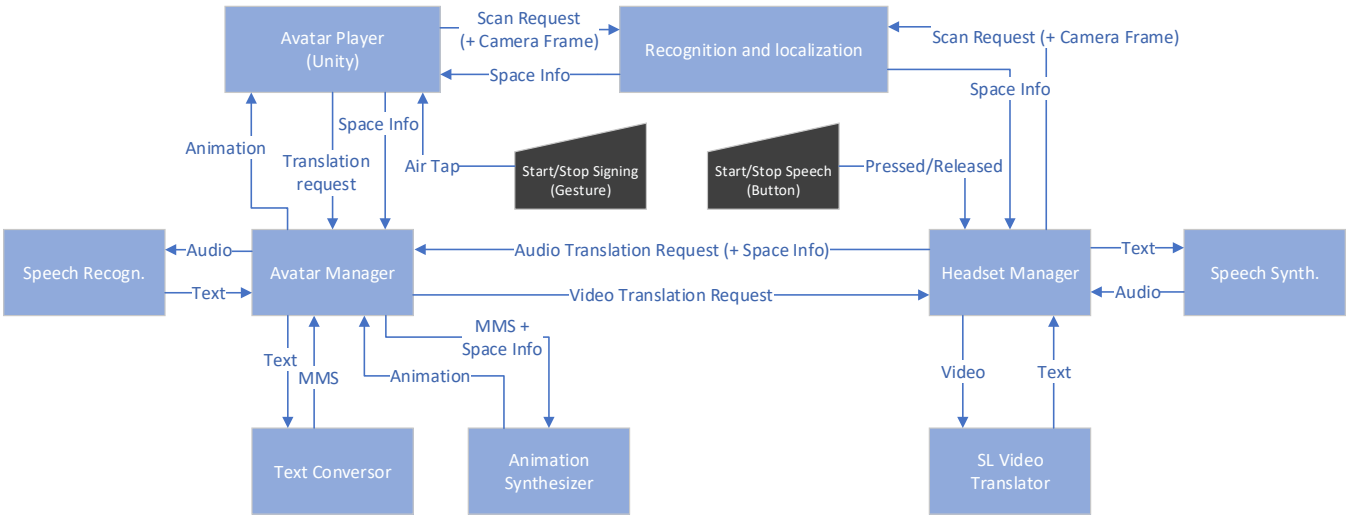

Fig. 2. Component diagram of the proposed application.

An MMS instance consists of a table listing the sequence of glosses to be displayed together with additional information about their composition (e.g., holding, parallel execution) and their inflection (e.g., relocation, eye brows modulation). It contains the necessary information needed to describe how signs can be inflected in order to refer to specific spatial locations. The detailed description of the MMS (already anticipated in [14]) is a work in progress.

When translating in the opposite direction, the Avatar Manager receives the translation request from the AR devices and forwards the request to the Headset Manager.

4.4. Avatar Player

The Avatar Player is the native application running on the AR device. It is designed to be as simple as possible, and its main duties are the recognition of special activation gestures (e.g., an air-tap) and forwarding information about the environment to the Avatar manager. Its most resource demanding operation is the animation of the avatar using the animation data incoming from the Avatar Manager.

Among the features of the Avatar Player there is for example the possibility to place the virtual interpreter on any horizontal surface of the surrounding environment and to scale the interpreter. This gives freedoms that go beyond the restrictions of the real world, such as the possibility to place a miniaturized version of the interpreter over a desk. Figure 3 shows a screenshot and a 3rd-person view of an alpha version of the App running on a mobile phone.

4.5. Text Conversor

Upon the request of the Avatar manager, this module produces an MMS instance from an input sentence. The conversion engine is an encoder–decoder model based on the transformer

Fig. 3. Preliminary version of the Avatar Player.

architecture [15]. While the encoder is a standard transformer encoder, to output a tabular MMS structure, we use multiple tied decoders to predict different types of outputs: glosses (textual), characteristics (boolean) and inflection parameters (real numbers).

4.6. Animation Synthesizer

This module takes as input an MMS instance and produces the animation data of the avatar, i.e., the animation curves describing the motion of the full-body skeleton and the face. This module (already partially implemented) consists of a set of Python scripts for the Blender 3D authoring tool. The current version is already able to parse MMS instances and compose on the timeline the animation concatenating the required glosses and inflecting hand trajectories and orientations. The development is now focusing on the inflection according to the Spatial Info, so that sign directions and eye-gaze are referring to the real location of signs in space.

Using Blender as a tool to generate animations has been already explored by Sharma and Filhol in synthesizing AZee utterances [16]. The advantage being the possibility to use a robust set of functions to analyze and manipulate animation data. The disadvantage is in the introduction of a significant delay between the request of animation and the resulting data, but our early tests shows that the realization of a 10-gloss sentence is executed in less than 10 seconds on a graphic workstation, with high margin for optimizations; a tolerable delay for a research prototype.

4.7. Headset Manager

The Headset manager is small software service running on a low-power microcontroller mounted the augmented headset, continuously sending images to the localization service and receiving an up-to-date status of the surrounding environment. In the voice-to-SL direction, the manager activates the microphone when the user presses its button, and sends the voice sample together with the space information to the Avatar Manager when the button is released.

In the opposite direction, when the headset receives a request for recognition, it activates its head-mounted camera, asks the Video Translator to convert the video snippet into text, and invoke the TTS service to play back the audio on the headphone speakers.

4.8. Sign Language Video Translator

This module consists of a visual features representation model and a textual model for the generation of spoken language sentences. The textual model is inspired by text-to-text transformer-based machine translation system. This architecture allows parallel processing of the whole sentence and facilitates the prediction on untrimmed video streams, following an approach that is widely used in recent works on SL translation ([17, 18]).

Similarly to the Text converter module, the performance of this architecture heavily depends on the amount of training data. We therefore propose a second architecture that gets rid of the need of annotated data (temporal boundaries in videos and glosses in text) [19] and uses pretrained sentence embeddings as additional supervision signal.

Another challenge of this module is the speed requirements of the real-time (RT) translation. Depending on the results of the first pilot experiments using all the connected devices, we will choose the best of our two approaches to ensure RT response together with translation quality.

4.9. Recognition and Localization

The goal of this software module is to receive a continuous stream of images of the surrounding environment, and return a map between a list of recognized objects and their 3D position in space. From the standpoint of computer vision, this would involve Visual Simultaneous Localization and Mapping (SLAM) and 3D Object Detection modules: a visual technique incrementally building a map of an unknown environment and estimate the 6DoF (position and orientation) pose of the camera with a strong focus on real-time operation [20].

For the proposed demonstrator here, 3D Reconstruction and pose estimation of cameras of both the participants will be realized by using Structure-PLP SLAM [21] as an out-of-the-box SLAM solution. It is a novel feature based SLAM system focused towards robust camera trajectory estimation and more structured mapping of the environment.

The module with be jointly working with one of the many existing state-of-the-art approaches and open-source frameworks for object category classification [22, 23], allowing to deliver the orientation and size of the objects relative to the frame of reference of the camera, therefore also obtaining the bounding box coordinates of objects with reference to the world co-ordinate frame.

5. CONCLUSION

We described the concept and the software architecture of a research project aiming at the implementation of a augmented reality virtual sign language interpreter that can simulate awareness of the surrounding environment and take advantage of spatial references when assisting in a dialog with a speaking partner.

We are aware that, at the present time, this is a kind of futuristic and ambitious project. The many components of the system are research prototypes tested in laboratory, simulated conditions, and their application in a real-scenario untested. Even the most stable components, as the 3rd party speech recognition system, are far from perfection and even more prone to errors when used in real noisy environments. The most critical part being the translation modules, which are current subject of intense research but still can't reach a fraction of the performances now expected for natural language processing.

Nevertheless, this will be a useful modular laboratory where each of the components can be replaced and tested in "in-the-wild" scenarios for their portability and generalization to noisy environments and unpredictable user behavior. Not to mention that this is a first attempt to provide real-time spatial information in the context of a dialogue for a sign language synthesis system. Our plan is to design a set of user studies to verify that the spatial awareness is, as easily expected, a desired feature, but also a motivation to investigate in a new direction on *how* such information can be embedded in a translation system. The grammatically correct way in which the avatar will then make use of spatial information is part of future work in collaboration with sign language linguists.

6. REFERENCES

[1] Lan Thao Nguyen, Florian Schicktanz, Aeneas Stankowski, and Eleftherios Avramidis, "Evaluating the translation of speech to virtually-performed sign language on AR glasses," in *13th Int. Conf. on Quality of Multimedia Experience (QoMEX)*, IEEE, Ed., 2021.

[2] Ralph Elliott, Javier Bueno, Richard Kennaway, and John Glauert, "Towards the Integration of Synthetic SL Animation with Avatars into Corpus Annotation Tools," in *LREC2010 Workshop on the Representation and Processing of Sign Languages*, Valletta, 2010, ELRA.

[3] Hernisa Kacorri and Matt Huenerfauth, "Implementation and Evaluation of Animation Controls Sufficient for Conveying ASL Facial Expressions," in *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, New York, 2014, ACM.

[4] John McDonald et al., "An automated technique for real-time production of lifelike animations of American Sign Language," *Universal Access in the Information Society*, vol. 15, no. 4, 2016.

[5] Larwan Berke, William Thies, and Danielle Bragg, "Chat in the Hat: A Portable Interpreter for Sign Language Users," in *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA, 2020, ACM.

[6] Nicoletta Adamo-Villani and Saikiran Anasingaraju, "Holographic signing avatars for deaf education," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 180, 2017, Springer Verlag.

[7] Antigoni Parmaxi and Alan A. Demetriou, "Augmented reality in language learning: A state-of-the-art review of 2014–2019," *Journal of Computer Assisted Learning*, vol. 36, no. 6, pp. 861–875, 2020.

[8] Le Luo et al., "Avatar Interpreter: Improving Classroom Experiences for Deaf and Hard-of-Hearing People Based on Augmented Reality," in *Extended Abstracts of the 2022 CHI Conference*, New York, USA, 2022, ACM.

[9] Karen Baldeon, William Oñate, and Gustavo Caiza, "Augmented Reality for Learning Sign Language Using Kinect Tool," in *Developments and Advances in Defense and Security*, Singapore, 2022, Springer.

[10] Marc Schröder and Jürgen Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," *International Journal of Speech Technology*, vol. 6, no. 4, 2003.

[11] Jonathan Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[12] Nickolay V. Shmyrev, "Vosk Speech Recognition Toolkit," https://github.com/alphacep/vosk-api.

[13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[14] Fabrizio Nunnari et al., "AVASAG: A German Sign Language Translation System for Public Services," in *1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. Aug. 2021, Association for Machine Translation in the Americas.

[15] Ashish Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[16] Paritosh Sharma and Michael Filhol, "Multi-Track Bottom-Up Synthesis from Non-Flattened AZee Scores," in *7th Workshop on Sign Language Translation and Avatar Technology*, Marseille, France, June 2022.

[17] Necati Cihan Camgöz et al., "Neural sign language translation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *IEEE/CVF CVPR*, 2020.

[19] Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet, "DFKI-MLT at WMT-SLT22: Spatio-temporal Sign Language Representation and Translation.," in *Seventh Conference on Machine Translation*, Abu Dhabi, UAE, 2022, ACL.

[20] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[21] Fangwen Shu, Jiaxuan Wang, Alain Pagani, and Didier Stricker, "Structure PLP-SLAM: Efficient sparse mapping and localization using point, line and plane for monocular, rgb-d and stereo cameras," *arXiv preprint arXiv:2207.06058*, 2022.

[22] Jean Lahoud and Bernard Ghanem, "2d-driven 3d object detection in rgb-d images," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[23] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.