Multi-modal Dialogue System with Sign Language Capabilities

M. Hrúz, P. Campr,
Z. Krňoul, M. Železný
Department of Cybernetics,
Faculty of Applied Sciences,
University of West Bohemia
Univerzitni 8, Pilsen 306 14,
Czech Republic
{mhruz, campr, zdkrnoul,
zelezny}@kky.zcu.cz

Oya Aran
Idiap Research Institute
Martigny, Switzerland
oya.aran@idiap.ch

Pınar Santemiz
Computer Engineering
Department, Boğaziçi
University
İstanbul, Turkey
pinar.santemiz
@bound.edu.tr

ABSTRACT

This paper presents the design of a multimodal sign-language -enabled dialogue system. Its functionality was tested on a prototype of an information kiosk for the deaf people providing information about train connections. We use an automatic computer-vision-based sign language recognition, automatic speech recognition and touchscreen as input modalities. The outputs are shown on a screen displaying 3D signing avatar and on a touchscreen displaying graphical user interface. The information kiosk can be used both by hearing users and deaf users in several languages. We focus on description of sign language input and output modality.

Categories and Subject Descriptors

H.4.3 [**Information Systems**]: Information systems applications—*Communications Applications*

General Terms

EXPERIMENTATION, LANGUAGES

Keywords

sign language, visual tracking, sign categorization

1. INTRODUCTION

Deaf and hearing impaired users have limited possibility of communication with hearing people. This can be a problem especially in the case when communicating with authorities or information providers. These people also cannot use speech-based automatic information services. In these cases dialogue systems should be designed to be accessible by deaf users. Our goal was to design such a dialogue system and verify its functionality on an information kiosk for deaf people. The system uses sign language (SL) as one of the communication means in both directions.

The hardware setup of the information kiosk (Fig. 1) is a standard PC, a touch screen display, a large screen, a microphone and several cameras. The cameras capture the body and facial gestures of the signing person, the large screen renders the SL output and the touch screen allows touch commands as an alternative input.

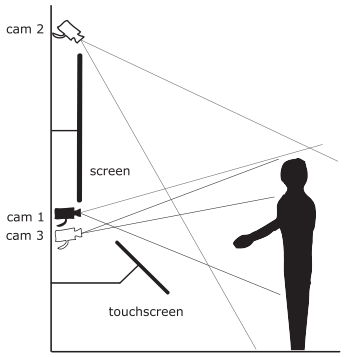

Figure 1: Setup of proposed information kiosk.

2. SIGN LANGUAGE RECOGNITION

The SL recognition system is intended for recognition of isolated signs. For this purpose a database was created using the described hardware setup. In total 338 files were recorded with one male and one female signer. The database contains 50 signs from Czech SL such as Czech towns, days etc. The constraining conditions were long sleeves, non-skin-colored clothes, uniform background and constant illumination. This is the first step towards an autonomous system. In the future much more data are required to train statistical models for recognition and less constraining conditions need to be applied. For tracking purposes we make use of joint particle filter that calculates a combined likelihood for all objects by modeling the likelihood of each object with respect to the others [1]. The filtering is done on a segmented image. We use skin-color segmentation to obtain the likelihood image. The result is shown in Fig 2.

The tracking provides coordinates of head and hands for each image from the camera. This information about trajectories is not sufficient for sign classification. In order to recognize the signs we also need shape information in addition to the tracking features. In order to describe the shape we have implemented five algorithms (3 based on Fourier

ASSETS'11, October 24–26, 2011, Dundee, Scotland, UK.
ACM 978-1-4503-0919-6/11/10.

(a) (b) (c)

Figure 2: (a) Original image, (b) Particle distribution with joint PF, (c) Estimated hand and head positions

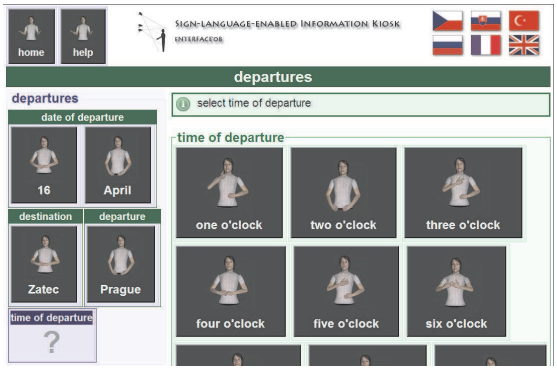

Figure 3: Sample screen of touchscreen graphical user interface.

descriptors, 1 based on Hu moments and last the one based on Discrete Cosine Transform, DCT). We tested their performances over the set of finger alphabet in Czech SL.

According to our results [1], DCT performed significantly better than the others. For the sign recognition from the calculated features we use Hidden Markov Model (HMM). The signs are modeled as an 8-state HMM (two of these states are non-emitting). Each state is modeled with a single Gaussian. This was due to the relative low amount of data. PCA and ICA were tested to reduce the dimensionality of the data and to align the data according to the feature space coordinate system. Both methods failed to improve the recognition rate since there were too few samples available. When more training data are available these methods should improve the recognition.

3. 3D SIGNING AVATAR

SL synthesis is used as the main output modality for the deaf users. We use the signing avatar animation in two forms. The first is an on-line generated avatar shown on the large screen that provides real-time feedback to deaf users. The second form is the pre-generated short movie clips inserted into the graphical user interface instead of text (see Fig. 3).

The animation model of the upper part of the human body is in compliance with the H-Anim standard. Currently, the animation model involves 38 joints and body segments. The talking head is composed from seven segments. The relevant body segments are connected by the avatar skeleton. For this purpose, one joint per segment is sufficient. The control of the skeleton is based on the rotation of segments (3 DOF per joint). The rotation of the shoulder, elbow, and wrist joints are computed by the inverse kinematics in accordance with 3D positions of the wrist and shoulder joints. The animation of the avatar's face, lips and tongue is rendered by the talking head system that performs local deformations of the relevant triangular surfaces.

For the manual component of the sign speech, the trajectory generator performs the syntactic analysis of HamNoSys symbolic strings. Since to define the rules and actions for all symbol combinations covering the entire notation variability is difficult, we have to make restrictions in order to preserve maximum degree of freedom. The trajectory generator currently involves 374 parsing rules. The structurally correct symbolic string is decomposed by the parsing rules to nodes of the parse tree. Two identical key frame data structures distinguishing the dominant and non dominant hand are used to describe the nodes. These data structures are composed from specially designed items. Firstly, the items of terminal nodes are filled from symbol descriptors stored in the definition file that currently covers 138 HamNoSys symbols.

Next, the nodes of the parse tree are processed by several tree walks whilst the rule actions are performed. We have defined 39 rule actions that are connected with each parse rule. The nodes are joined together and transformed to the control trajectories. The final trajectories of both hands are generated in the root node. The final step is the concatenation that puts together trajectories of hands with the articulatory trajectories generated by the talking head system and provides the control over a signed utterance. Reader can find more details in [2].

4. CONCLUSION

We have presented a dialogue system with SL capabilities. The system is able to track isolated signs and provide features for the recognizer. Significantly more data are needed for successful recognition. The system can be trained with signs from any topic. The system provides feedback in the form of a signing avatar. The dialogue is computer driven and suitable for different scenarios where the user answers queries with isolated signs.

5. ACKNOWLEDGMENTS

This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET 101470416, by the Grant Agency of the Czech Republic, project No. GAČR 102/09/P609 and by the Ministry of Education of the Czech Republic, project No. ME08106.

6. REFERENCES

[1] P. Campr, M. Hrúz, A. Karpov, P. Santemiz, M. Železný, and O. Aran. Sign-language-enabled information kiosk. In *eNTERFACE'08*, 2009.

[2] Z. Krňoul, J. Kanis, M. Železný, and L. Müller. Czech text-to-sign speech synthesizer. In *Proceedings of the 4th international conference on Machine learning for multimodal interaction*, MLMI'07, pages 180–191, Berlin, Heidelberg, 2008. Springer-Verlag.