Multimodal Assisted Living Environment

Principal Investigators: Lale Akarun, Alexey Karpov, Hulya Yalcin

Candidate Participants: Yunus Emre Kara, Haşim Sak

Abstract: The aim of this project is to establish a multimodal environment for an assisted living smart space. The system will make it possible for users to access the assisted living information system through easy-to-use and less complicated multimodal techniques such as speech, hand and facial gesture recognition. Using different modalities, we would like to communicate through speech, gesture, head/eye tracking as well as combining several modalities to form a multimodal interface that can interpret the requirements of the user. This multimodal interface will receive the input in whatever way the user is capable of conveying input and receive output whatever form user is capable of perceiving. For instance, if the user has a difficulty of hearing and speaking, visual information will be used. If the person has limited mobility but has perfect verbal communication and hearing abilities, auditory channels will be used. A combination of modes will be used to meet the requirements of circumstances with multiple disabilities.

Project objectives

The project aims to incorporate ubiquity to the user interface through the use of following technologies; the first set of technologies are microphone and camera arrays. Using microphone arrays will allow speech recognition at a distance and using camera arrays, we aim to extend visual capabilities of the elderly artificially after processing the video sequences by state-of-art computer vision techniques. Nevertheless, for practical reasons it is not viable to install too many cameras.

In order to address daily living tasks, convenience scenarios need to be designed. We will develop application scenarios for multimodal interaction with the assisted living information system as well as emergency scenarios such as a fall. In both regular and emergency cases, the user may need to communicate with the information system remotely using speech and gestures. We will develop multimodal human-computer interaction techniques that are able to work in these scenarios.

Different spoken languages such as English, Russian and Turkish may be considered for the project depending on the composition of the project team.

The objectives of the project are the following:

- Designing scenarios and interfaces for a close to real time multimodal information system.
- Designing various modules of the system that is required to complete the given task.
    - Continuous Speech recognition module
    - Speech Synthesis (TTS) module
    - Human Activity Detection Module
    - Recognition of Orientation and Pointing Gestures
    - Facial Expression Detection Module

Background information

Human Activity Detection:

Vision based human action recognition is the process of assigning action labels to videos. In action recognition, videos are readily segmented in time. A video is expected to have only one action. Additionally, an activity is defined to be an ordered set of actions. For example, cooking is an activity whereas stirring is an action. Human action understanding has many application areas concerning security, surveillance, assisted living, and even entertainment. Access control, person identification, anomaly detection, and human-computer interaction are some of the areas that can benefit from human motion analysis. In activity detection, first, the activity is localized over time and space and the label assignment follows.

The feature extraction methods from the images of a video sequence can be divided into two categories: global and local representations [Poppe2010]. In global representations, the person is localized first, then the region of interest is encoded as a whole. On the other hand, in local representations, some patches in the images are encoded separately. These patches are usually extracted from the neighborhoods of interest points. The representation is found by combining the information of these patches.

Speech Recognition:

Human's speech refers to the processes associated with the production and perception of sounds used in spoken language, and automatic speech recognition (ASR) is a process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a software or hardware

module. Several kinds of speech are identified: spelled speech (with pauses between phonemes), isolated speech (with pauses between words), continuous speech (when a speaker does not make any pauses between words) and spontaneous natural speech. The most common classification of ASR by recognition vocabulary is following [Rabiner93]:

- small vocabulary (10-1000 words);
- medium vocabulary (up to 10 000 words);
- large vocabulary (up to 100 000 words);
- extra large vocabulary (up to and above million of words that is adequate for inflective or agglutinative languages)

Recent automatic speech recognizers exploit mathematical techniques such as Hidden Markov Models (HMMs), Artificial Neural Networks (ANN), Bayesian Networks or Dynamic Time Warping (dynamic programming) methods. The most popular ASR models apply speaker-independent speech recognition though in some cases (for instance, personalized systems that have to recognize owner only) speaker-dependant systems are more adequate.
In framework of the given project a ASR system will be constructed using the Hidden Markov Model Toolkit (HTK version 3.4) [Young06]. Language models based on statistical text analysis and/or finite-state grammars will be implemented for ASR [Rabiner08].

Speech Synthesis:
Speech synthesis is the artificial production of human speech. Speech synthesis (also called text-to-speech (TTS) system converts normal orthographic text into speech translating symbolic linguistic representations like phonetic transcriptions into speech. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database (compilative speech synthesis or unit selection methods) [Dutoit09]. Systems differ in the size of the stored speech units; a system that stores allophones or diphones provides acceptable speech quality but the systems that are based on unit selection methods provide a higher level of speech intelligibility. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create voice output. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood (intelligibility).

Detailed technical description

a) Technical description

The project has the following work packages

WP1. Design of the overall system

In this work package, the design of the overall system will be implemented. The system will be operating close to real-time and will take input from speech recognition, human activity detection and perform designated actions through graphical outputs and synthesized speech outputs.

WP2. Human Activity Detection

Human Activity Detection will be implemented for predetermined number of activities.

WP3. Speech Recognition

Continuous speech recognition will be implemented. Language models will be used to solve ambiguities.

WP4. Speech Synthesis

Speech synthesis will be implemented.

WP5. System Integration and Module testing

The modules implemented in WP2-WP5 will be tested and integrated in the system designed in WP1.

WP6. Integrated System Interface Demos

The Systems integrated in WP5 will be tested extensively and demo videos will be recorded.

b) Resources needed: facility, equipment, software, staff etc.

- The training databases for the recognition tasks should be ready before the project. Additional data will be collected for adaptation and test purposes.
- Prototypes or frameworks for each module should be ready before the starting date of the project. These preparations are essential successful completion of the project, since the project duration is short
- High resolution cameras with high frame-per-second rate are needed for human activity detection and facial expression recognition.
- A computer dedicated to the demo application is required.
- Staff with enough expertise is required to implement each of the tasks mentioned in the detailed technical description.
- C/C++ programming languages will be used for implementation.

c) Project management

Each participant will have a clear task that is parallel with their expertise.
Required camera hardware will be provided by the leaders.

d) Work plan and implementation schedule

A tentative timetable detailing the work to be done during the workshop;

	Week 1	Week 2	Week 3	Week 4
WP1. Design of the overall system				
WP2. Human Activity Detection				
WP3. Speech recognition				
WP4. Speech Synthesis				
WP5. System Integration and Module testing				
WP6. Integrated System Interface Demos				
Documentation				

Benefits of the research

The deliverables of the project will be the following:
D1: Human activity detection module
D2: Facial Expression Recognition module
D3: Speech Recognition module
D4: Speech Synthesis module
D5: Integrated system interfaces
D6: Final Project Report

Profile of team

Leaders

Short CV - Lale Akarun

Lale Akarun is a professor of Computer Engineering in Bogazici University. Her research interests are face recognition and HCI. She has been a member of the FP6 projects Biosecure and SIMILAR, COST 2101: Biometrics for identity documents and smart cards, and FP7 FIRESENSE. She currectly has a joint project with Karlsruhe University on use of gestures in emergency management environments, and with University of Saint Petersburg on Info Kiosk for the Handicapped. She has actively participated in eNTERFACE workshops, leading projects in eNTERFACE06 and eNTERFACE07, and organizing eNTERFACE07.

Selected Papers:

- Pinar Santemiz, Oya Aran, Murat Saraclar and Lale Akarun , Automatic Sign Segmentation from Continuous Signing via Multiple Sequence Alignment, Proc. IEEE Int. Workshop on Human-Computer Interaction, Oct. 4, 2009, Kyoto, Japan.
- Oya Aran, Lale Akarun, "A Multi-class Classification Strategy for Fisher Scores: Application to Signer Independent Sign Language Recognition, Pattern Recognition, accepted for publication.
- Cem Keskin, Lale Akarun, " Input-output HMM based 3D hand gesture recognition and spotting for generic applications", Pattern Recognition Letters, vol. 30, no. 12, pp. 1086-1095, September 2009.
- Oya Aran, M.S. Thomas Burger, Alice Caplier, Lale Akarun, "A Belief-Based Sequential Fusion Approach for Fusing Manual and Non-Manual Signs", Pattern Recognition, vol.42 no.5, pp. 812-822, May 2009.
- Oya Aran, Ismail Ari, Alexandre Benoit, Pavel Campr, Ana Huerta Carrillo, Franois-Xavier Fanard, Lale Akarun, Alice Caplier, Michele Rombaut, and Bulent Sankur, "Signtutor: An Interactive System for Sign Language Tutoring". IEEE Multimedia, Volume: 16 Issue: 1 Pages: 81-93, Jan-March 2009.
- Oya Aran, Ismail Ari, Pavel Campr, Erinc Dikici, Marek Hruz, Siddika Parlak, Lale Akarun & Murat Saraclar, Speech and Sliding Text Aided Sign Retrieval from Hearing Impaired Sign News Videos , *Journal on Multimodal User Interfaces*, vol. 2, n. 1, Springer, 2008.
- Arman Savran, Nese Alyuz, Hamdi Dibeklioğlu, Oya Celiktutan, Berk Gokberk, Bulent Sankur, Lale Akarun: "Bosphorus Database for 3D Face Analysis", The First COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008), Roskilde, Denmark, 7-9 May 2008.
- Alice Caplier, Sébastien Stillittano, Oya Aran, Lale Akarun, Gérard Bailly, Denis Beautemps, Nouredine Aboutabit & Thomas Burger, Image and video for hearing impaired people, *EURASIP Journal on Image and Video Processing*, Special Issue on Image and Video Processing for Disability, 2007.

Former eNTERFACE projects:

- Aran, O., Ari, I., Benoit, A., Carrillo, A.H., Fanard, F., Campr, P., Akarun, L., Caplier, A., Rombaut, M. & Sankur, B, "SignTutor: An Interactive Sign Language Tutoring Tool", Proceedings of eNTERFACE 2006, The Summer Workshop on Multimodal Interfaces, Dubrovnik, Croatia, 2006.
- Savvas Argyropoulos, Konstantinos Moustakas, Alexey A. Karpov, Oya Aran, Dimitrios Tzovaras, Thanos Tsakiris, Giovanna Varni, Byungjun Kwon, "A multimodal framework for the communication of the disabled", Proceedings of eNTERFACE 2007, The Summer Workshop on Multimodal Interfaces, Istanbul, Turkey, 2007.
- Ferda Ofli, Cristian Canton-Ferrer, Yasemin Demir, Koray Balcı, Joelle Tilmanne, Elif Bozkurt, Idil Kızoglu, Yucel Yemez, Engin Erzin, A. Murat Tekalp, Lale Akarun, A. Tanju Erdem, "Audio-driven human body motion analysis and synthesis", Proceedings of eNTERFACE 2007, The Summer Workshop on Multimodal Interfaces, Istanbul, Turkey, 2007.
- Arman Savran, Oya Celiktutan, Aydın Akyol, Jana Trojanova, Hamdi Dibeklioglu, Semih Esenlik, Nesli Bozkurt, Cem Demirkır, Erdem Akagunduz, Kerem Calıskan, Nese Alyuz, Bulent Sankur, Ilkay Ulusoy, Lale Akarun, Tevfik Metin Sezgin, "3D face recognition performance under adversarial conditions", Proceedings of eNTERFACE 2007, The Summer Workshop on Multimodal Interfaces, Istanbul, Turkey, 2007.

Short CV – Alexey Karpov

Alexey Karpov received his MSc from St. Petersburg State University of Airspace Instrumentation and PhD degree in computer science from St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), in 2002 and 2007, respectively. His main research interests are automatic Russian speech and speaker recognition, text-to-speech systems, multimodal interfaces based on speech and gestures, audio-visual speech processing, sign language synthesis. Currently he is a senior researcher of Speech and Multimodal Interfaces Laboratory of SPIIRAS. He has been the (co)author of more than 80 papers in refereed journals and International conferences, for instance, Interspeech, Eusipco, TSD, etc. His main research results are published by the Journal of Multimodal User Interfaces and by the Pattern Recognition and Image Analysis (Springer). He is a coauthor of the book "Speech and Multimodal Interfaces" (2006), and a chapter in the book "Multimodal User Interfaces: From Signals to Interaction" (2008, Springer). He leads several research projects funded by Russian scientific foundations. He is the winner of the 2-nd Low Cost Multimodal Interfaces Software (Loco Mummy) Contest. Dr. Karpov is a member of organizing committee of series of the International conferences "Speech and Computer" SPECOM, as well as member of the EURASIP and ISCA. He took part at eNTERFACE workshops in 2005, 2007 and 2008.

Researchers needed:

- MS or PhD students with good C/C++ programming knowledge to work on the individual modules and integration.

References

[Dutoit09] Dutoit T., Bozkurt B. Speech Synthesis, Chapter in Handbook of Signal Processing Acoustics, D. Havelock, S. Kuwano, M. Vorländer, eds. NY: Springer. Vol 1, pp. 557-585, 2009.

[Laptev2005] Laptev, I. (2005). On Space-Time Interest Points. International Journal of Computer Vision, 64(2-3), 107-123. doi: 10.1007/s11263-005-1838-7.

[Laptev2008] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (p. 1–8). IEEE.

[Messing2009] Messing, R., Pal, C., & Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. IEEE 12th International Conference on Computer Vision, 104-111. Ieee. doi: 10.1109/ICCV.2009.5459154.

[Moeslund2006] Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 104(2-3), 90-126. doi: 10.1016/j.cviu.2006.08.002.

[Poppe2007] Poppe, R. (2007). Vision-based human motion analysis: An overview. Computer Vision and Image Understanding, 108(1-2), 4-18. doi: 10.1016/j.cviu.2006.10.016.

[Poppe2010] Poppe, R. (2010). A survey on vision-based human action recognition. Image and Vision Computing, 28(6), 976-990. doi: 10.1016/j.imavis.2009.11.014.

[Rabiner93] Rabiner L., Juang. Fundamentals of Speech Recognition New Jersey: Prentice-Hall, Englewood Cliffs, 1993.

[Rabiner08] Rabiner L., Juang B. Speech Recognition, Chapter in Springer Handbook of Speech Processing (Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng, eds.), NY: Springer,
2008.

[Young06] Young S. et al. The HTK book version 3.4 Manual. Cambridge University Engineering Department, Cambridge, UK, 2006