Proc. IROS-2015 Workshop on "Cognitive Mobility Assistance Robots",
Hamburg, Germany, Sep. 2015.
http://robotics.ntua.gr/IROS2015-Workshop-Cognitive-Mobility-Assistance/

MULTIMODAL SENSORY PROCESSING FOR HUMAN ACTION RECOGNITION IN MOBILITY ASSISTIVE ROBOTICS

P. Maragos, V. Pitsikalis, A. Katsamanis, N. Kardaris, E. Mavroudi, I. Rodomagoulakis and A. Tsiami

School of ECE, National Technical University of Athens, 15773 Athens, Greece
{maragos,vpitsik,nkatsam,irodoma,antsiami}@cs.ntua.gr

1. INTRODUCTION

One of the main objectives of the EU project MOBOT [1], which generally aims at the development of an intelligent active mobility assistance robot, is to provide multimodal sensory processing capabilities for human action recognition. Specifically, a reliable multimodal information processing and action recognition system needs to be developed, that will detect, analyze and recognize the human user actions based on the captured multimodal sensory signals and with a reasonable level of accuracy and detail within the context of the MOBOT framework for intelligent assistive robotics. Different sensory modalities need to be combined into an integrated human action recognition system. One of the main thrusts in the above effort is the development of robust and effective computer vision techniques to achieve the visual processing goals based on multiple cues such as spatio-temporal RGB appearance data as well as depth data from Kinect sensors. Another major challenge is the integration of recognizing specific verbal and gestural commands in the considered human-robot interaction context.

In this presentation we summarize advancements in three tasks of the above multimodal processing system for human-robot interaction (HRI): action recognition, gesture recognition and spoken command recognition.

2. ACTION RECOGNITION

Our approach to detect and classify human actions from continuous RGB-D video streams, captured by visual sensors on the MOBOT robotic platform, consists of the following main steps: visual feature extraction, feature pre-processing and encoding, and the classification. An initial baseline version of our system was based on detecting space-time interest points, computing descriptors in a neighborhood around these points [e.g. Histogram Of Gradient (HOG) [3], Histogram of Flow (HOF), and HOG3D], using the Bag-of-Features (BoF) representation of the videos, and classification with Support Vector Machines (SVMs); such systems have exhibited promising performance in movie action classification [6]. Subsequently, we have enriched several sub-components of this pipeline by developing state-of-the-art approaches, as explained in [2]. Specifically, for the visual features we employ approaches such as spatio-temporal interest points by computing spatio-temporal energies via our multiscale Gabor 3D detector [7] on the RGB or Depth visual streams, as well as dense trajectories [10]. Then several descriptors capture appearance and motion information. State-of-the-art encoding methods employed include i) vector quantization and ii) vector of locally aggregated descriptors [4]. After feature encoding we train discriminative classifiers, such as SVMs, and classify a video segment containing a single action instance by employing different state-of-the-art variants of the widely used bag of visual words framework. In our set of tools employed (either in post-processing or in gesture recognition), we also combine SVMs with Hidden Markov Models (HMMs) and related algorithms. Overall, our system automatically detects human activity, classifies detected actions and localizes them in time; see Figure 1 for an overview of the system's pipeline. All the above have been evaluated on both the MOBOT dataset as well as on known datasets found in the literature. Our recognition results reach 86% on the MOBOT dataset and 93% on the KTH dataset. Details can be found in [2].

3. GESTURE RECOGNITION

Gesture recognition concerns the communication of the elderly subjects with the platform via a predefined set of gestural commands. There are several challenges faced during our work with the MOBOT dataset. For instance, it is usual to have alternative pronunciations of the same gesture among performances by different users. Further, in the MOBOT case, mobility disabilities seriously impede the performance ability of a gesture for some users, and therefore, alternative pronunciations are more frequent. Our gesture recognition systems shares some methodologies with the visual action recognition system. Initially, for the visual processing we used the RGB video stream, combined with pose information that became available after pose annotation. The extracted features included either handshape or movement information. For handshape features we focused on a neighbourhood of

This research work was supported by the European Union under the project MOBOT with grant FP7-ICT-2011.2.1-600796.

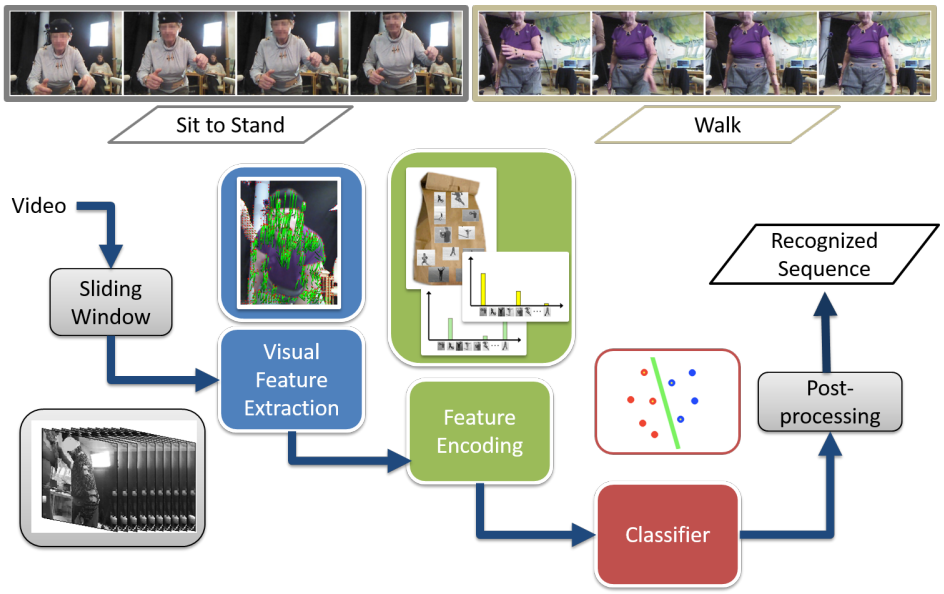

Fig. 1: Visual action recognition system overview. Top: Actions performed by patients in the MOBOT dataset. Bottom: Action localization and classification pipeline.

the hand centroid, so that we can use local descriptors such as HOG to extract features on the handshape. For movement-position features we used the available pose annotation to compute characteristics about position and motion of the arms (positions, velocities, accelerations of hands and elbows mostly). We have evaluated the complete framework of feature extraction and gesture learning based on HMMs for the statistical modeling. Our experimental results on the 2013 ACM Gesture Challenge dataset can found in [8] and preliminary results for the MOBOT data set can be found in [2]. More recently, in an effort to view gestures as refined visual actions, we have developed a visual front-end for gesture recognition that is based on the same approach used for action recognition, i.e. dense trajectories, feature encoding, and SVMs. This newer approach on gesture data showed that we can get roughly similar results to the ones obtained with our previous system, but without employing any manual (human provided) pose annotations. Our current gesture recognition systems has an average performance of about 70% on the MOBOT dataset, by using only motion-appearance features extracted from the RGB data. Our ongoing plans include the incorporation of an automatic pose annotation system.

4. SPOKEN COMMAND RECOGNITION

In the context of multimodal processing for human action recognition, we have developed a first version of an online system for always-listening spoken command recognition in German that is integrated on the ROS-based robotic platform and operates with an 8-channel MEMS microphone array. Based on the multichannel input, the module is designed to detect and recognize the user's intention to execute a specific operation of the robotic assistant. For instance, the elderly user may call the system by uttering a keyword like "MOBOT" and then provide a voice command from a predefined set of commands that are included on the recognition grammar, e.g, "MOBOT, turn right". The detection and recognition tasks are expected to be challenging due to the distant speaking configuration which is prone to noise and reverberation effects depending on the acoustic environment in which the session is taking place. Additional challenges may be introduced due to the existence of background speech and non-speech events possibly overlapping with the keyword and command segments to be detected and recognized. An overview of the implemented multichannel speech processing pipeline is depicted in Fig. 3. To support always-listening operation, the pipeline is built on the widely used cascade of three speech processing stages: a) voice activity detection, to separate speech from non-speech events, b) key-phrase detection based on the keyword-filler approach, to identify a predefined system activation phrase, and c) grammar-based automatic speech recognition, to recognize the issued command. All stages are applied to the denoised signal derived after delay and sum beamforming of the MEMS channels. Context-dependent German triphones have been trained on 55 hours of publicly available read speech and used for keyword spotting and recognition. Promising results were obtained after testing the system on MOBOT data. Two tests were conducted: i) the first on 8 patients seated approximately two meters in front of the robotic platform providing verbal and non-verbal (gestural) commands and ii) the second on 10 normal German-speaking users which held and followed the platform operat-

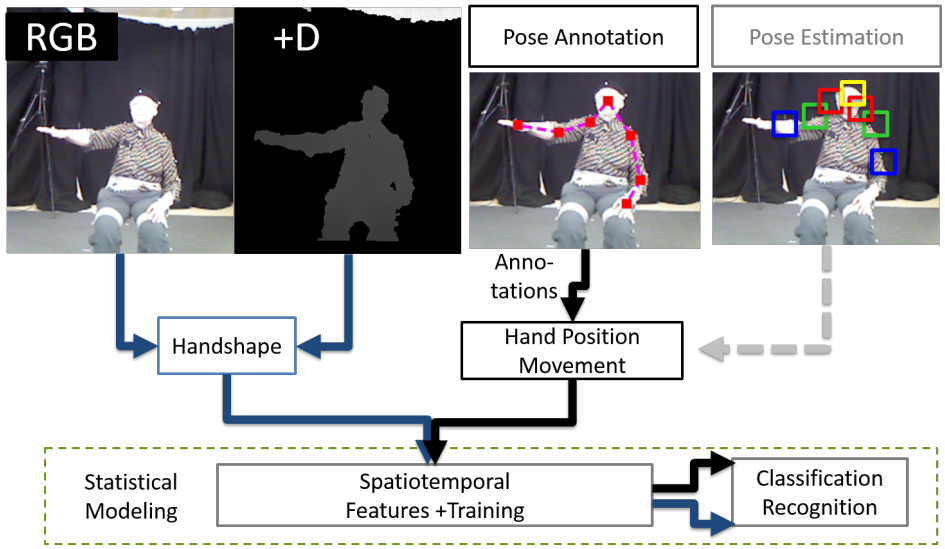

Fig. 2: Overview: Visual gesture recognition. Multiple information channels are combined within a common framework.

ing in a "following mode". The achieved average word accuracies of 73% and 85% on leave-one-out experiments (testing on one speaker after global MLLR adaptation of the acoustic models to the other speakers) renders the system usable as stand-alone or combined with the other modalities. More details about the employed methods for key-word spotting and recognition can be found in our previous work [5].

5. MULTIMODAL SENSOR FUSION

Within the MOBOT objective of multisensory processing for HRI, we have also been working with the design and experimentation of fusion algorithms for the integration of gestural and spoken command recognition. Such a cross-modal integration can significantly increase performance. Our first experimental system was based on a multimodal sensor fusion for audio-visual gesture recognition that exploited the color, depth and audio information captured by a Kinect sensor. Recognition of a time sequence of audio-visual gesture commands was based on an optimized fusion of all different cues and modalities (audio, movement-position, handshape). Our system [8, 9] was evaluated on the ACM 2013 Gesture Challenge dataset where it outperformed all other competing published approaches and achieved a 93% accuracy. We are currently adapting this multimodal action-gesture-speech recognition system for the MOBOT dataset and are developing a real-time version on the ROS robotic platform.

Acknowledgement: We wish to thank Kevis Maninis and George Pavlakos for their contributions to action recognition and gesture recognition while working at the NTUA IRAL, and Antonis Arvanitakis for his help with the ROS platform.

6. REFERENCES

[1] MOBOT: Intelligent active MObility assistance roBOT integrating multimodal sensory processing, proactive autonomy and adaptive interaction. *EU project*. `http://www.mobot-project.eu/`

[2] "MOBOT Deliverable D1.2: Intermediate Report on Multimodal Human Action Recognition, Work Package 1: Multimodal Sensory Processing for Human Action Recognition", Mar. 2015.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2005)*, Jun. 2005, vol. 1, pp. 886–893.

[4] H. Jegou, M. Douze, C. Schmid, and P. Perez,"Aggregating local descriptors into a compact image representation", *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2010)*, Jun. 2010, pp. 3304–3311.

[5] A. Katsamanis, I. Rodomagoulakis, G. Potamianos, P. Maragos and A. Tsiami, "Robust far-field spoken command recognition for home automation combining adaptation and multichannel processing", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP 2014)*, pp. 5547–5551, Florence, Italy, May 2014.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies", *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2008)*, Jun. 2008, pp. 1–8.

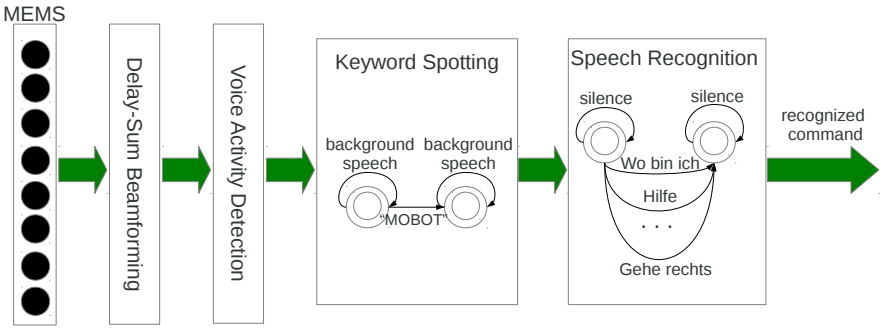

Fig. 3: An overview of the always-listening spoken command recognition pipeline with finite-state-automaton (FSA) representations of the finite state grammars employed for keyword spotting and recognition.

[7] K. Maninis, P. Koutras and P. Maragos,"Advances on Action Recognition in Videos Using an Interest Point Detector Based on Multiband Spatio-Temporal Energies", *Proc. IEEE Int'l Conf. Image Processing (ICIP 2014)*, Paris, France, Oct. 2014.

[8] G. Pavlakos, S. Theodorakis, V. Pitsikalis, A. Katsamanis and P. Maragos, "Kinect-Based Multimodal Gesture Recognition Using a Two-Pass Fusion Scheme", *Proc. IEEE Int'l Conf. Image Processing (ICIP 2014)*, Paris, France, Oct. 2014.

[9] V. Pitsikalis, A. Katsamanis, S. Theodorakis and P. Maragos, "Multimodal Gesture Recognition via Multiple Hypotheses Rescoring", *Journal of Machine Learning Research*, vol.16, pp.255–284, Feb. 2015.

[10] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories", *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2011)*, Jun. 2011, pp. 3169–3176.