Multimodal Signal Processing and Learning Aspects of Human-Robot Interaction for an Assistive Bathing Robot

A. Zlatintsi, I. Rodomagoulakis, P. Koutras, A.C. Dometios, V. Pitsikalis, C.S. Tzafestas and P. Maragos

ICCS, School of ECE, National Technical University of Athens 15773, Greece

http://robotics.ntua.gr
http://cvsp.cs.ntua.gr

1. Outline - Contributions

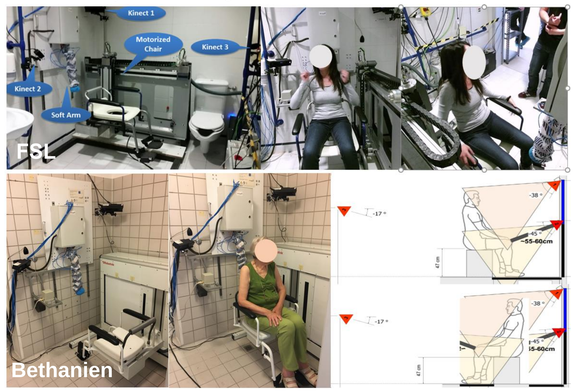
- Explore new aspects of assistive living on smart human-robot interaction (HRI) involving automatic recognition and online validation of speech and gestures in a natural interface.
- New experimental framework and resources of a real-life scenario for elderly users, within the framework of I-SUPPORT bathing robot, addressing health and hygiene care issues.
- New domain specific dataset and a suite of tools used for data acquisition and a state-of-the-art pipeline for multimodal learning with emphasis on audio and RGB-D visual streams, considering also privacy issues by evaluating the depth visual stream, using Kinect sensors.

Goal:

- Development of a robotic bathing system that will assist towards independent living and improved life quality for elderly.
- Enhancement the communication making it natural, intuitive and easy to use, thus, enhancing it with respect to social aspects.

2. I-Support: Robotic Shower System

- ***Core system functionalities*:** tasks for bathing the legs and the back (taking into account impairments, limitations and user requirements).
- ***Multi-view system architecture***
  - 3 Kinect sensors for 3D pose reconstruction of the scene (user and robot) and identification of user gestures;
  - Audio system including 8 distributed condenser microphones



I-Support automatic bathing environment. Setup of the Kinect sensors as installed in the two validation sites: FSL (top) and Bethanien (bottom) hospitals.

Audio-gestural data acquisition

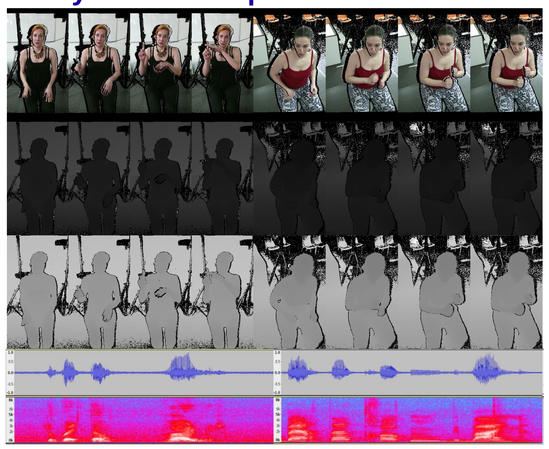
.. for modeling the user-robot interaction

Core bathing tasks: washing/scrubbing/wiping/rinsing the back or legs changing base settings of the system, i.e., temperature, water flow etc. spontaneous/emergency commands and a background model

- ***Visual data***: 23 users performing predefined gestures
- ***Audio data***: 8 users uttering predefined spoken commands (German)

Total number of commands for each task:
- 25 and 27 gesture commands for washing the legs and the back, resp.
- 23 spoken commands – preceded by a keyword ("Roberta")

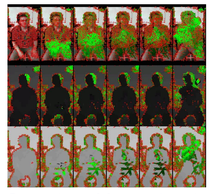
3. System description



Data streams acquired by sensors #1 and #2: RGB (top), depth (2nd row) and log-depth (3rd row) frames from a selection of gestures ("Temperature Up", "Scrub Legs"), accompanied by the corresponding German spoken commands – waveforms (4th row) and spectrograms (bottom row) – "Wärmer", "Wasch die Beine".

Visual processing for system architecture

- Emploing **Dense Trajectories** features along with a) Bag-of-Visual-Words (BoVW) framework and b) VLAD (Vector of Locally Aggregated Descriptors) encoding, using SVM.
- The Dense Trajectories method main concept consists in sampling feature points from each video frame on a regular grid and tracking them through time based on optical flow.



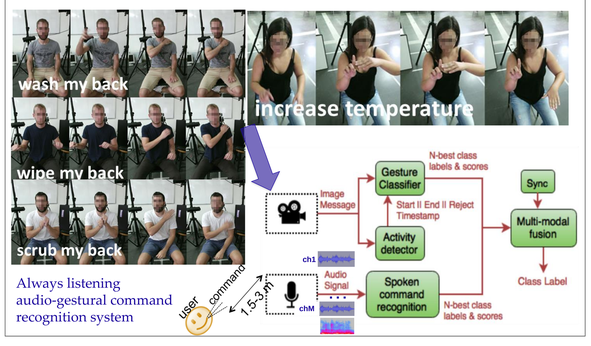
Employed descriptors:
Trajectory descriptor, HOG, HOF and Motion Boundary Histograms (MBH).

A non-linear transformation of depth using logarithm (log-depth) enhances edges related to hand movements leading to richer dense trajectories on the ROI, close to the result obtained using the RGB stream.

Comparison of dense trajectories extraction over the RGB (top), depth (middle) and log-depth (bottom) clips of gesture "Scrub Back".

Audio processing for system architecture

- Robustness of the system via:
  a) denoising of the far-field multi-channel signals
  b) global MLLR adaptation of the acoustic models and
  c) combined command detection/recognition.



Always listening audio-gestural command recognition system

4. Online validation of the online A-G system with primary users

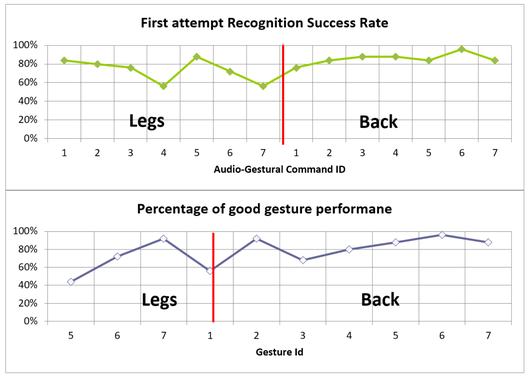
- Evaluation of the system's functionality and the HRI between the I-Support bathing robot and the primary end-users (elderly), using audio and audio-gestural commands.
- ***Two simulated bathing scenarios***: bathing the legs and back at dry conditions.
- ***Two pilot sites***:
  - the Fondazione Santa Lucia (FSL) Hospital (Rome, Italy)
  - the Bethanien Hospital (Heidelberg, Germany).
- ***Statistically significant number of users***: 25 and 29 patients (having various cognitive impairments).
- Results delivered as ROS messages to the system's finite state machine (FSM) to:
  a) decide the action to be taken after each recognized command,
  b) control the various modules and
  c) manage the dialogue flow by producing the right audio feedback to the user.
- ***Challenges***: acoustic noise, low voice, weak/inaccurate motions/gestures, small variations in the camera setup, no additional data collection from the primary users

Vocabulary		
English	**Italian**	**German**
Wash legs	Lava le gambe	Wasch meine Beine
Wash back	Lava la schiena	Wasch meinen Rücken
Scrub back	Strofina la schiena	Trockne meinen Rücken
Stop (pause)	Basta	Stop
Repeat (continue)	Ripeti	Noch einmal
Halt	Fermati subito	Wir sind fertig

Audio-gestural commands incl. in the two scenarios, preceded by the keyword "Roberta".

	Distal Region		Back Region	
ID	**Command**	**Modality**	**Command**	**Modality**
1	Wash Legs	A	Wash Back	A-G
2	Stop	A	Halt	A-G
3	Repeat	A	Scrub Back	A-G
4	Halt	A	Stop	A-G
5	Wash Legs	A-G	Repeat	A-G
6	Halt	A-G	Halt	A-G
7	Halt	A-G	Halt	A-G

Experimental protocol exhibiting a variety of 7Audio (A) and/or Audio-Gestural (A-G) commands. Sequence of the commands as performed by the participants.



Recognition statistics per command (FSL data). The vertical red lines distinguish the command sequences (see the corresponding IDs) for the tasks "washing the legs" and "washing the back".

	System Performance %						User Performance %			
	MCRR %			Accuracy %			Speech		Gestures	
	L	B	Av.	L	B	Av.	L	B	L	B
FSL	80	87	83.5	86	73	79.5	98	99	81	78
Bethanien	85	74	79.5	67	77	72	91	90	84	71

Multimodal recognition evaluated in terms of 1) Multimodal Command Recognition Rate (MCRR): MCRR= # of commands correctly recognized by the system / # of commands correctly performed by the user, 2) accuracy (%) and 3) user performance/learning rate (%).
Average Audio-Gestural Command Recognition Results; system performance (%) and user performance (%) averaged across 25 and 29 users at FSL and Bethanien Hospitals, respectively.

5. Offline experimental results

Audio classification

Feat.	Task: Legs	Task: Back
MFCC	75.8	67.6

Average (leave-one-out) classification results for the pre-defined spoken commands, performed by 8 subjects for the two bathing tasks.

Gesture classification

		Task: Legs		Task: Back	
Feat.	Encoding	RGB	D	RGB	D
Traj.	BoVW	69.64	60.52	77.84	60.87
HOG	BoVW	41.01	53.34	58.51	57.14
HOF	BoVW	74.15	66.26	82.92	71.58
MBH	BoVW	77.36	65.31	80.81	65.73
Comb.	BoVW	**80.88**	**74.41**	**83.92**	**75.70**
Traj.	VLAD	69.22	52.66	74.34	54.14
HOG	VLAD	49.86	65.99	61.23	65.63
HOF	VLAD	76.54	72.88	83.17	78.07
MBH	VLAD	78.35	75.12	82.54	73.09
Comb.	VLAD	**83.00**	**78.49**	**84.54**	**81.18**

Average (leave-one-out) classification accuracy (%) results for the pre-defined gestures performed by 23 subjects. Results for the four features and their combination (Comb.), using the two encodings are shown for RGB data and D (depth) data for the two bathing tasks.

6. Conclusions

- We presented a multimodal interface for an assistive bathing robot and a real-life use case, providing a rich set of tools and data.
- Such resources can be employed to develop natural interfaces for multimodal interaction between humans and robotic agents.
- We further investigate how the communication between end-users and the robot will be as intuitive and effortless as possible using co-speech gesturing, which is the most natural way for human-human communication, while also enhancing the recognition, in cases of speech dysfluencies or kinetic problems.
- The online results are really promising given the difficulties of the task.
- By sharing such resources, we aim to build a public crowdsourced library that shall open new perspectives in smart assistive HRI.
- Available software:
  http://robotics.ntua.gr/projects/building-multimodal-interfaces

References

[1] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *Proc. ICASSP* 2016.
[2] J. Werle and K. Hauer, "Design of a bath robot system-user definition and user requirements based on international classification of function- ing disability and health (ICF)," in *Proc. RO-MAN* 2016.
[3] N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis, and P. Maragos, "A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands," in *Proc. ACM* 2016.
[4] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Found. trends human-computer Interact.*, vol. 1, no. 3, pp. 203– 275, Feb. 2007.
[5] R. Kachouie, S. Sedighadeli, R. Khosla, and M.-T. Chu, "Socially assistive robots in elderly care: A mixed-method systematic literature review," *Int'l Jour. Human-Computer Interaction*, vol. 30, no. 5, pp. 369–393, May 2014.
[6] F. Rudzicz, R. Wang, M. Begum, and A. Mihailidis, "Speech interaction with personal assistive robots supporting aging at home for individuals with alzheimer's disease," *ACM Trans. Access. Comput.*, vol. 7, no. 2, pp. 1–22, July 2015.
[7] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE ICCV* 2013.

Acknowledgments

This research work was supported by the EU under the project I- SUPPORT with grant H2020-643666.
More information can be found at: http://www.i-support-project.eu/
Contact: [nzlat, maragos]@cs.ntua.gr