Proceedings of IEEE Int'l Conference on Image Processing (ICIP-2016), Phoenix, AZ, USA, Sep. 2016.

A MULTIMEDIA GESTURE DATASET FOR HUMAN ROBOT COMMUNICATION: ACQUISITION, TOOLS AND RECOGNITION RESULTS

I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, A. Arvanitakis, P. Maragos

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

{irodoma,vpitsik,maragos}@cs.ntua.gr, {nick.kardaris, antonisar}@gmail.com

ABSTRACT

Motivated by the recent advances in human-robot interaction we present a new dataset, a suite of tools to handle it and state-of-the-art work on visual gestures and audio commands recognition. The dataset has been collected with an integrated annotation and acquisition web-interface that facilitates on-the-way temporal ground-truths for fast acquisition. The dataset includes gesture instances in which the subjects are not in strict setup positions, and contains multiple scenarios, not restricted to a single static configuration. We accompany it by a valuable suite of tools as the practical interface to acquire audio-visual data in the robotic operating system, a state-of-the-art learning pipeline to train visual gesture and audio command models, and an online gesture recognition system. Finally, we include a rich evaluation of the dataset providing rich and insightfull experimental recognition results.

Index Terms— human-robot communication, multimedia gesture dataset, visual gesture recognition, audio commands

1. INTRODUCTION

None would refuse the impact of newly published datasets in human activity recognition tasks. By focusing specifically on audio-visual gesture recognition, we notice an emerging trend of new approaches, competition challenges [1, 2], together with a few datasets that have been recently acquired [1, 2, 3]. Nevertheless, for the sub-field of human gesture recognition, there is a lack of datasets [4]; this is in contrast to broader related fields, such as unconstrained action videos, sports, movies and human activities [5, 6, 7], to name but a few. Our contributions concern a rich suite of tools that assists data acquisition of new audio-visual gestures, a training pipeline, and an online system for visual and audio gesture commands recognition. The impact of such resources would be high, if we consider that the tools are not restrictive to the specific assistive HRI task.

Indicative existing datasets, related to the area of multimodal, audio-visual gesture recognition include cases such as the MSRC-12 [3], the recent ChaLearn 2014 [8], and others [2, 9, 10, 11, 12, 13]. Main parameters in such datasets include the potential multimodality, the sensor's type and point of view, the user position, the number of gestures, and the ground-truth annotations. Most of the datasets include a single front-view, a medium size vocabulary (10-30 classes), a single fixed position of the subjects, either standing or sitting, except from [14] where the subject is moving. When it comes to the sensor, most of them nowadays employ the kinect sensor. For instance the MSRC-12 [3] is captured with Kinect, includes 12 classes and 30 subjects. Chalearn contains 20 classes of Italian cultural/anthropological signs; The dataset of [14] includes 14 classes with three subjects for a total of 126 sequences captured from a moving camera and in complex backgrounds.

The introduced dataset offers several differential contributions. We tackle a crucial aspect of variability so as to have more realistic conditions. This concerns the non-constant position of the subject, and then, to a smaller degree, the angle-of-view. By introducing such variance as well as by letting the subjects perform the gestures more freelly and from a not-strict position, we manage to introduce variability which would benefit of model generalization. Finally, there is the multimedia aspect of the data, which apart from the Kinect, includes a microphone array for audio.

The grund-truth annotations are important, in terms of their production and their employment by the learning algorithms. Similar to scripted-ground-truth [4], we apply what we term as "ground-truths by-construction". This allows us to accomplish fast dataset acquisition: i.e. acquired by a reusable acquisition interface, *at once*, together with its accompanying annotations, which are in general hard to get. Additionally, the acquisition interface supports the dataset extension with new gestures, out-of-vocabulary or background cases.

At the same time the dataset is accompanied by a suite of tools, to train audio and visual models, as well as to recognize in a online manner gesture utterances. These tools are based on the current state-of-the-art including dense trajectory descriptors [15, 16], bag-of-visual-words encoding, and support vector machines [17]. These allow us to provide: (1) a framework that can be re-used either as is, or by varying components such as the feature extractor, or the encoding; (2) trained models, and (3) a recognizer employing these models that lets someone test audio-visual gestures from uknown subjects. Finally, together with the audio and visual classification experiments, we elaborate on valuable practical questions such as how many subjects are required to reach a specific accuracy percentage given the included audio and visual variability.

2. THE HUMAN-ROBOT COMMUNICATION DATASET

Motivation: The purpose of this dataset is the continuation of our previous work on multimodal human-robot interaction (HRI) [18] focusing on the challenging task of mobility assistive robotics for elderly people. Our research has been mainly conducted within the European project MOBOT [1] where the goal is to implement a prototype in order to enable communication of an impaired user with the robotic platform. To achieve robust and real-time recognition, we extent the multimodal MOBOT database [19] with the current dataset which has been designed to enable the exploration of several parameters of our system as described in the remaining text.

This research work was supported by the European Union under the project MOBOT with grant FP7-ICT-2011.2.1-600796 and in part under the EU project I-SUPPORT with grant H2020-643666.

[1] www.mobot-project.eu

Fig. 1: Left: Standing gesture "no" at position A; right: Command "come here" at the same position, darker environment.

Fig. 2: Up: Standing near (position A), "What time is it"; bottom: sitting far (position B), "I want to perform a task".

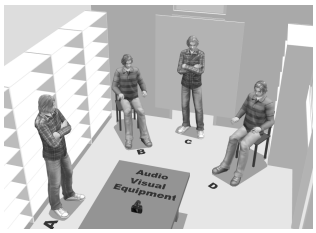

Fig. 3: Experimental setup

Loc.	Camera	MEMS
A	-/90	-/62
B	224/173	180/155
C	-/165	-/99
D	190/140	164/125

Fig. 4: Distances (in cm) from locations (Loc.) A, B, C, D to the sensors, as depicted in Fig. 3

The acquired dataset includes video and multichannel audio data. These are to be employed in automatic visual gesture recognition and in automatic speech recognition. The subjects were in front of the kinect in various setups, following prerecorded video prompts. The data acquisition framework is implemented in the Robotic Operating System (ROS) that is used to collect the data while annotating on-the-way the gesture time boundaries with a web-based GUI. The data are available in both ROS-bags and video[2].

2.1. Experimental setup and dataset description

The subjects were allowed to position themselves in an approximate manner near to markers so as to introduce variance to our dataset. The marker distances to the audio-visual equipment are as listed in Fig. 4. The acquired dataset includes 1438 commands as executed by 13 subjects. These samples are audio-visual spoken phrases and gestures from a vocabulary of 19 commands [3], and two actions "Sit-to-Stand" and "Stand-to-Sit". Each command has been recorded on average 74 times. The sequence of the gestures is randomized per subject so as not to include identical sequences, or identical iterations in the row.

[2]The dataset as well as several tools e.g. to visualize it can be found in `http://robotics.ntua.gr/datasets/ROSAudioVideo`, reviewing user-name and password: "mobotdataset".

[3]Vocabulary list: "Avoid Obstacle", "Come Closer", "Come Here", "Go Away", "Go Straight", "Go Through Door", "Help", "Lets Go", "Perform Task", "Stop", "Turn Left", "Turn Right", "I Want to Sit Down", "I Want to Stand Up", "Where Am I" "What time is it", "Yes", "No", "Park".

Fig. 5: Top: standing at medium distance (position C) "Avoid obstacle"; middle: Sit-to-Stand at position D; bottom: sitting at medium distance (position B) "Come Near".

Scenario	1A	1B	1C	2A	2A	2A	2B	2D
Dist.	near	med	med	near	near	near	far	med
Pose	Stand	Sit	Stand	Stand	Stand	Stand	Mix	Mix
Angle	Right	Front	Front	Right	Left	Front	Front	Front
Inst.	152	302	285	95	95	95	210	204

Table 1: Partitioning of acquisition scenarios wrt the acquisition parameters and gesture instances (Inst.).

2.2. Acquisition Scenarios

The dataset includes multiple setups, that introduce variability in terms of the parameters discussed bellow: (1) the standing or sitting pose of the subjects; (2) the background; (3) the angle of view of the camera (4) the distance of the subject to the recording equipment.

In the first case the subjects are standing, as the example shown in Fig. 1. At random intervals of the recording the lighting configuration is updated, as shown in the same figure. For the sitting case, the subject is sitting in front of the equipment while the acquisition includes multiple setups, as described in Table 1.

Mixed sitting-standing poses (Mix): During the corresponding recording sessions, sit-to-stand and stand-to-sit actions are also recorded in between gestures. These are important in the context of the involved HRI task, and fit to the rest of the vocabulary. For instance, when the subject utters a command "I want to stand up", while the subject is sitting, then the next natural action to be performed is an actual "Sit-to-Stand", and the rest of the commands are articulated at the stand pose. In this way the subject's pose changes from sitting to standing and vice versa within the same recording.

Angle, Distance, Background: In one of the setups, we vary the camera angle of view. The standard setup is defined, unless speci-

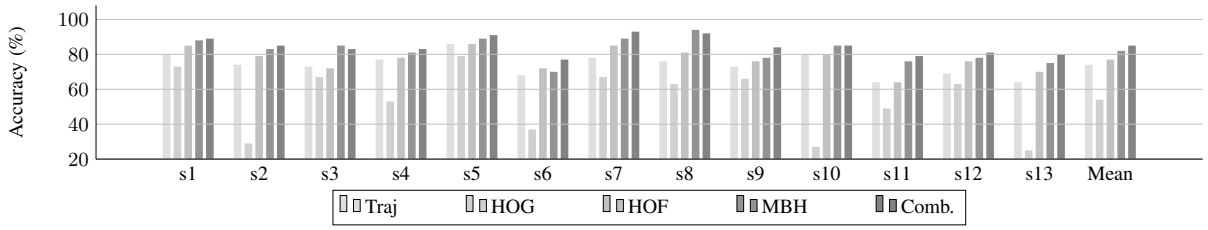

Fig. 6: Visual gesture accuracy of all descriptors, per subject, and mean accuracy for all subjects.

fied, as the front. Two tilted versions are introduced by letting the subjects to face the camera in left and right at 10 degree angles. The distance of the recording equipment is also varied according to the positions specified in Table 4 and Fig. 3. Note finally, that due to the different setups the background is greatly updated.

Parameters and subjects: One of the differences with existing datasets consists of the variability introduced in the dataset due to the mixed setups even for a single subject. Subjects 1–8 took part in the setups 1A, 1B, 1C; subjects 9–13 took part in the setups 2A, 2B, 2D. The second character A, B, C, corresponds to the positions as depicted in Fig. 3. The rest of the varying parameter combination are shown in Table. 1. Samples of the data collected in various setups are shown in Fig. 1, Fig. 2 and Fig. 5.

3. RECOGNITION EXPERIMENTS AND RESULTS

3.1. Visual modality

Feature extraction and descriptors: Dense trajectories [15] consists in sampling feature points from each video frame on a regular grid and tracking them through time based on optical flow. Tracking is performed in multiple spatial scales, and trajectories are pruned to a fixed length L to avoid drifting. Following the trajectory extraction, different descriptors are computed within space-time volumes along each trajectory[4].

Feature encoding: Extracted features are encoded using separate codebooks per descriptor. Codebooks are constructed by clustering a subset of selected training features into K clusters. Each trajectory is assigned to its closest visual word. We use Bag of Visual Words (BoVW) encoding, i.e. a histogram of visual word occurrences, yielding a sparse K-dimensional video representation.

Classification and fusion: Videos are classified based on their BoVW representation, employing non-linear support vector machines (SVMs) with the χ^2 kernel [20]. Descriptors are combined in a multichannel approach, by computing distances between BoVW histograms as: $K\left(\mathbf{h}_i, \mathbf{h}_j\right) = \exp\left(-\sum_c \frac{1}{A_c} D\left(\mathbf{h}_i^c, \mathbf{h}_j^c\right)\right)$, where c is the c-th channel, i.e. $\mathbf{h}_i^c$ is the BoVW representation of the i-th video, computed for the c-th descriptor, and A_c is the mean value of χ^2 distances $D\left(\mathbf{h}_i^c, \mathbf{h}_j^c\right)$ between all pairs of training samples. Since we face multiclass classification problems, we follow the one-against-all approach and select the class with the highest score.

3.2. Audio modality

Speech modeling and recognition: Speaker independent acoustic models are trained for Greek based on the Logotypografia corpus [22]. HTK tools and recipe [23] was followed for training 3-state, cross-word triphones, with 8 Gaussians per state, based on MFCC-plus-derivatives. The recognizer for spoken commands is grammar-based: this is a robust solution in small tasks, like the examined HRI task. We also denoise by delay-and-sum beamforming of 7 MEMs[5] [24] channels arranged on a 4 cm linear setup. The Greek triphones are adapted to a development set consisting of recordings from the corpus. We use the global maximum likelihood linear regression (MLLR) technique, to transform the means of the Gaussians yielding improvements in distant spoken command recognition [25].

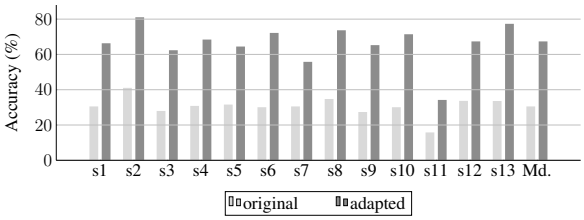

Fig. 7: Spoken command classification accuracy on the full set of 19 audio-gestural commands using the original Greek triphones and the adapted version; the median (Md.) is also depicted.

3.3. Recognition results

Experimental setup: Experiments are carried out in a leave-one-out scheme, i.e. testing on single subjects, while the rest are used for training. This is repeated for all subjects giving an unbiased estimate on the ability to generalize to unseen subjects. We report classification accuracy for each subject individually, as well as the mean accuracy over all.

Visual gestures recognition: In Fig. 6 we assess the performance of all descriptors, as well as their combination (see Sec. 3.1). Motion-based descriptors consistently perform best, since motion, rather that shape or appearance, is the most discriminative cue for our visual vocabulary. Nevertheless, in most cases combining all descriptors results in better accuracy.

Audio command phrases recognition: As shown in Fig. 7 the performance is approximately doubled after adapting the original acoustic models to the specific conditions of the dataset and to the particularities of speech signals after denoising with beamforming. Nevertheless, the average performance of 67% across the testing subjects shows that the task is challenging: diffused background noise exists on some sessions, speech is far-field, and subjects are uttering from various positions and angles towards the MEMS. Additionally, although the recognition grammar is designed to support

[4]Descriptors include: the Trajectory descriptor for motion, HOG [20] for shape and appearance, and finally, for motion, HOF [21] and MBH [15] computed on both axes (MBHx, MBHy).

[5]Micro Electro-Mechanical System (MEMS) microphones are compact sensors with comparable performance to the traditional electret condenser microphones (ECMs).

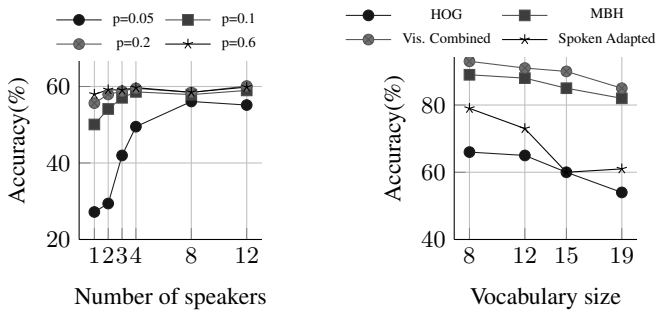

Fig. 8: Mean accuracy varying the number of adaptation speakers and the percentage p of randomly selected samples from each one.

Fig. 9: Accuracy of gesture and spoken commands depending on the vocabulary size.

acoustically distinguishable commands, we observed that commands with small length of one word are very often misrecognized.

Varying the vocabulary and training subjects subset: Figure 9 depicts how the vocabulary size affects the accuracy. Several audio-visual commands share common motion or audio sub-units, thus increasing the vocabulary size has a negative impact on the discriminative ability. A valuable question that arises often for a dataset or a practical recognition system, is how many data are enough to get an acceptable accuracy. Of course one may just use as much as possible, although this may come at a cost. Next, we provide some evidence on this. We train models on subsets of subjects, by sampling the subjects' combinations space. To avoid any bias, we randomly sample them in order to contain $n = [1, 2, 4, 6, 8, 12]$ subjects; that is take for instance multiple (35) training sets of $n = 4$ subjects. This number of instances is proportional, e.g. 10% to the number of possible combinations. Figure 10 shows the distribution of the recognition accuracy over the whole set of experiments that are carried out with the same number of n. The bottom row depicts the shift between consecutive histograms measured by their χ^2 distance normalized by the corresponding increment of n.

First, for the audio, the histograms on the left column of Fig. 10 show that by increasing the number of subjects used in adaptation, the average performance per subject increases but non-linearly, as shown in the bottom row of Fig. 10: the accuracy distributions tend to shift unequally for each increment of n. Another factor concerns the percentage p of adaptation samples (sentences) used for each subject included in the adaptation sets. Figure 8 depicts the average performance across the testing subjects when varying n and p. Global MLLR adaptation reaches quickly its maximum performance by using $p = 10\%$ of the samples from $n = 3$ subjects which is similar to using $n = 12$ subjects with $p = 60\%$ of their samples.

For the visual case, the shifts of histograms on the right column of Fig. 10 are more clear. The χ^2 distances in the bottom indicate that the performance increment is more noticeable when training on $n = [1 \dots 3]$ subjects and stabilizes for $n > 3$. Nevertheless, training on $n = 8$ seems to be sufficient and robust enough compared to the scenario of using $n = 12$ subjects.

System generalization and extensibility: The employed acquisition and experimentation framework allows us to easily extent our data and test our system in a variety of conditions and setups. Experimentation on this new dataset shed light on how the performance depends on parameters such as the number of training/adaptation subjects and the vocabulary size. Previous results on the MOBOT database showed the challenges that exist in very realistic scenarios

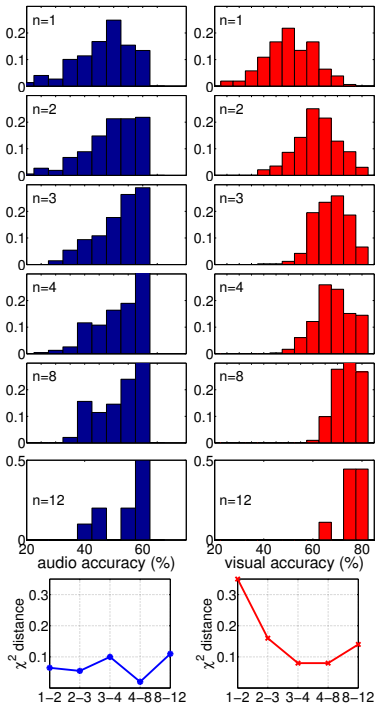

Fig. 10: Testing the effectiveness of training of the visual, audio models by varying the number of employed subjects (n). The histograms show how the accuracies per testing subject distribute when using samples from $n = [1, 2, 3, 4, 8, 12]$ subjects for training and testing to unseen sets consisting of the rest subjects. The shift between consecutive histograms is quantified using the χ^2 distance normalized by the increment of n.

with elderly subjects. Currently, based on the presented framework, we continue collecting data from user studies on elderly subjects in order to increase our system's robustness.

4. CONCLUSION

In this work we presented a new multimedia dataset for audio and visual gesture-commands for human-robot-interaction. This has enough variability in terms of setup and acquisition scenarios such as sitting vs standing, lighting variations, variable distances and angles of the acquisition equipment. We accompany the dataset with a rich set of tools such as the web-controlled acquisition interface, that can be used to introduce by-construction ground-truths, to acquire additional data, or out-of-vocabulary words. The backbone of our suite of tools, is the visual and audio pipeline for training the models, and the online recognizer. Apart from the rich recognition results for both modalities, we show evidence in a standard experimenter's question: "how many subjects and data are required?". We discovered that while employing state-of-the-art processing pipelines three subjects are enough to adapt the employed acoustic models in audio modality and eight subjects to train from scratch the employed SVMs in visual modality. The experiments are to be further enriched by multimodal fusion but this is out of the scope of this article. We expect that work in focused datasets, which are sparsely existing, shall advance the state-of-the-art in automatic visual gesture, and multimedia recognition, as well as its application to human-robot-interaction.

5. REFERENCES

[1] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proc. of the 15th ACM on Int'l Conf. on Multimodal Interaction*. ACM, 2013, pp. 445–452.

[2] S. Ruffieux, D. Lalanne, and E. Mugellini, "ChAirGest: A Challenge for Multimodal Mid-air Gesture Recognition for Close HCI," in *Proc. of the 15th ACM Int'l Conf. on Multimodal Interaction*, New York, NY, USA, 2013, ICMI '13, pp. 483–488, ACM.

[3] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. 2012, CHI '12, pp. 1737–1746, ACM, http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/.

[4] S. Ruffieux, D. Lalanne, E. Mugellini, and O. A. Khaled, "Gesture recognition corpora and tools: A scripted ground truthing method," *Computer Vision and Image Understanding*, vol. 131, pp. 72–87, 2015.

[5] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proc. Int'l Conf. on Comp. Vis.*, 2011, pp. 2556–2563.

[6] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[7] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, 2009, pp. 2929–2936.

[8] S. Escalera, X. Bar, J. Gonzlez, M. Bautista, M. Madadi, M. Reyes, V. Ponce-Lpez, H. Escalante, J. Shotton, and I. Guyon, "ChaLearn looking at people challenge 2014: Dataset and results," in *Computer Vision-ECCV 2014 Workshops*, vol. 8925, pp. 459–473. 2015.

[9] L. Liu and L. Shao, "Learning discriminative representations from RGB-D video data," in *Proc. of the 23th Int'l Joint Conf. on Artificial Intelligence*, 2013, pp. 1493–1500.

[10] V. Bloom, D. Makris, and V. Argyriou, "G3d: A gaming action dataset and real time action recognition evaluation framework," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 7–12.

[11] A. Sadeghipour, L.-P. Morency, and S. Kopp, "Gesture-based object recognition using histograms of guiding strokes," in *Proc. Brit. Machine Vis. Conf.*, 2012.

[12] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 2012, pp. 1737–1746.

[13] A. Just, O. Bernier, and S. Marcel, "HMM and IOHMM for the recognition of mono-and bi-manual 3D hand gestures.," in *BMVC*, 2004, pp. 1–10.

[14] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. Int'l Conf. on Comp. Vis.*, 2009, pp. 444–451.

[15] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, 2011, pp. 3169–3176.

[16] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. Int'l Conf. on Comp. Vis.*, 2013, pp. 3551–3558.

[17] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *arXiv preprint arXiv:1405.4506*, 2014.

[18] 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. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, 2016, pp. 2702–2706.

[19] E.-S. Fotinea, E. Efthimiou, A.-L. Dimou, T. Goulas, P. Karioris, A. Peer, P. Maragos, C. Tzafestas, I. Kokkinos, K. Hauer, et al., "Data acquisition towards defining a multimodal interaction model for human–assistive robot communication," in *Universal Access in HCI. Aging and Assistive Environments*, pp. 613–624. Springer, 2014.

[20] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Machine Vis. Conf.* BMVA Press, 2009, pp. 124–1.

[21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, 2008, pp. 1–8.

[22] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, "Large vocabulary continuous speech recognition in Greek: Corpus and an automatic dictation system," in *Proc. Interspeech*, 2003.

[23] S. J. Young et al., *The HTK Book, (for HTK version 3.4)*, Cambridge University Engineering Department, 2006.

[24] Z. Skordilis, A. Tsiami, P. Maragos, G. Potamianos, L. Spelgatti, and R. Sannino, "Multichannel speech enhancement using mems microphones," in *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, 2015, pp. 2729–2733.

[25] 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," in *Proc. Int'l Conf. on Acoustics, Speech and Sig. Proc.*, 2014, pp. 5547–5551.