MULTIMODAL HUMAN ACTION RECOGNITION IN ASSISTIVE HUMAN-ROBOT INTERACTION

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

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

{irodoma,vpitsik,nkatsam,antsiami,maragos}@cs.ntua.gr, nick.kardaris@gmail.com, emavrou1@jhmi.edu

ABSTRACT

Within the context of assistive robotics we develop an intelligent interface that provides multimodal sensory processing capabilities for human action recognition. Human action is considered in multimodal terms, containing inputs such as audio from microphone arrays, and visual inputs from high definition and depth cameras. Exploring state-of-the-art approaches from automatic speech recognition, and visual action recognition, we multimodally recognize actions and commands. By fusing the unimodal information streams, we obtain the optimum multimodal hypothesis which is to be further exploited by the active mobility assistance robot in the framework of the MOBOT EU research project. Evidence from recognition experiments shows that by integrating multiple sensors and modalities, we increase multimodal recognition performance in the newly acquired challenging dataset involving elderly people while interacting with the assistive robot.

Index Terms— multimodal sensor processing, assistive robotics, speech recognition, action-gesture recognition

1. INTRODUCTION

Human actions are inherently multimodal. Their recognition is a multilevel problem since they include audio-visual cues posing challenges at the level of features, information stream modeling and fusion. Nevertheless, multimodal human action recognition, which is currently viewed under multiple viewpoints in the literature, is still considered an open research field. Related application areas show great variability: human-computer interaction [1], multimedia indexing and retrieval, surveillance and multimodal event detection. In all the above, human actions and related events have been mainly studied *only* with respect to the individual modalities, apart from a few exceptions. This lack becomes more apparent when the actions' multimodal nature complicates things, e.g., when audio events and spoken commands, body actions, hand gestures and interactions between multiple subjects are involved. Such multimodal actions are of great value in human-computer interaction, and especially in assistive robotics. Our aim is to account for multiple modalities in human action recognition, while focusing, but not constraining ourselves, in a challenging *assistive human-robot interaction* (HRI) task for elderly people employing a newly acquired dataset [2].

"Human action" is usually translated in the literature as visual action [3]. This is the reason that most state-of-the-art approaches deal only with visual cues [4] and the most popular datasets [5] are considered with respect to visual cues [6]. Recently, several works address issues raised by multiple modalities. The most common aspect is multimedia event detection as in [7, 8, 9]. Another viewpoint, appeared in a recent challenge [10] concerning aspects of audio/visual cues for multimodal voice-gestures [11, 12, 13, 14]. Other multimodal applications list gesture and accompanying speech integration [15], integration for agreement recognition [16], social signal analysis [17], and complex events [18].

Multimodal human action recognition poses challenges due to distant speech recognition and noise, pronunciation variability, scene noise by other subjects, camera motion and variation in action/gesture performance. Another source of difficulties concerns the nature of our task, i.e. elderly subjects who often articulate or pronounce the multimodal gestures in a loose manner. Overcoming such problems for each modality separately is still open. Further, there are issues related to fusion despite numerous efforts in this direction [19], since getting satisfactory unimodal results does not necessarily guarantee better multimodal results. Last, application in HRI poses additional challenges: (a) either practical such as the real-time computation, integration issues in the robotic platform, (b) or research, concerning other involved modalities, e.g., laser range data, and the specific needs of the elderly. In this context, multimodal assistive HRI works are only recent and sparse [20, 21, 22].

To fill this gap, we introduce multimodal action recognition[1] in assistive HRI for elderly subjects. We deal with the individual modalities by advancing and adapting state-of-the-art approaches in automatic speech recognition as well as in visual action recognition. In the former, we employ beamformed multi-microphone array processing for robust distant speech recognition, whereas in the latter we adopt the state-of-the-art approach based on dense trajectories. Then, we late fuse the separate information streams to obtain the multimodal result. The approaches are evaluated on a challenging newly acquired HRI assistive task [2]. For the individual modalities we provide supplementary evaluations [6] showing that the employed approaches can be successfully applied in a broader context. Finally, we briefly describe how practically this multimodal assistive HRI is exploited in the MOBOT robotic platform providing intuition on our goal of assistive robotics, as well as information on the robotic platform integration for real-time performance.

In previous works, we investigated multichannel distant speech recognition [23] for domestic environments. The employed methods are further explored within this work on other datasets and are integrated into an online system for the robotic assistant. For multimodal fusion we proposed a more generic scheme [14] applied in the ChaLearn dataset [10].

This research work was supported by the EU under the project MOBOT (www.mobot-project.eu) with grant FP7-ICT-2011.2.1-600796 and in part under the project I-SUPPORT with grant H2020-643666.

[1] We generally employ the term "recognition" which is currently implemented in the online system (see the concluding remarks in Sec. 7). For the quantitative experiments (Sec. 6) presented we show *classification accuracies*, that is, after employing loose segmentation boundaries. Nevertheless, this does not affect the generality of our approach.

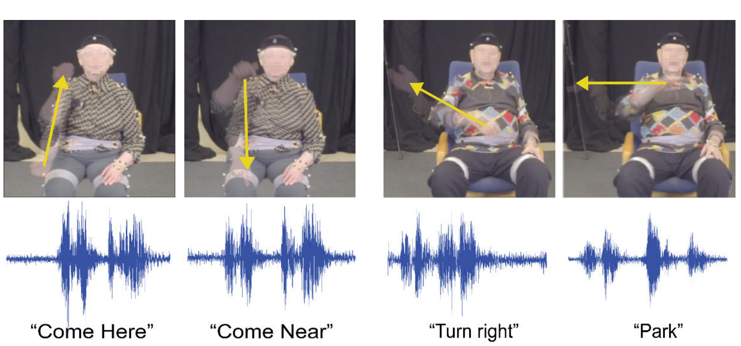

Fig. 1: Sample gestural and audio commands from the MOBOT 6.a dataset. The two leftmost pictures show two commands that are acoustically very similar but their corresponding gestures are distinguishable. On the other end, the last two pictures show commands whose gestural performances are similar but differ a lot acoustically.

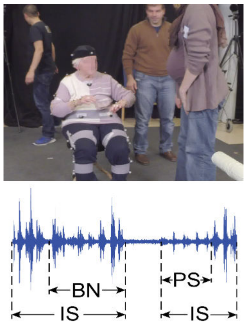

Fig. 2: Challenges related to the MOBOT-6a dataset. IS: Instructor speaking, PS: Patient speaking, BN: Background Noise

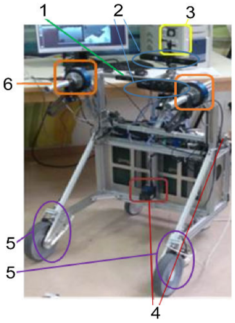

Fig. 3: MOBOT prototype rollator: 1. MEMS array, 2. Kinect 3. GoPro camera, 4. laser range finders, 5. encoders, 6. force/torque sensors

2. HRI ASSISTIVE TASK MULTIMODAL DATASET

The MOBOT multimodal database consists of multiple tasks based on the daily life of elderly people with mobility disabilities.

MOBOT robotic platform: The experimental prototype used for data acquisition consists of a passive rollator equipped with sensors, as shown in Fig. 3, such as: laser range sensors scanning the walking area for environment mapping and obstacle detection, and at the back detecting lower limbs movement; force/torque handle sensors, and visual sensors: a HD camera to record patient's upper body movements and two Kinect sensors. One Kinect captures the torso, waist and hips and the second faces downwards at the lower limbs. Finally, an array of 8-microphone MEMS is used for audio capturing.

Multimodal data: Herein we take advantage of the HD camera and the microphone array inputs. The MOBOT-6a task includes 19 different gestural and verbal commands developed to accommodate the communication with the robotic platform. The patient is sitting in front of the rollator, placed at a distance of 2.5 meters. Each command is performed by the 13 patients, $3-6$ times.

Challenges: Mobility disabilities seriously impede the performance ability of a verbal and/or gestural command for some users, and therefore, alternative pronunciations are frequent and diverse. Due to the cognitive disabilities of some users, in some cases we observe different pronunciations of a command even among multiple performances of the same user.

3. AUDIO GESTURE COMMANDS

Speech modeling and recognition: Speaker independent acoustic models are trained for German on 55 hours of publicly available close-talked clean read speech from adult speakers. The HTK tools and recipe [24] was followed for training 3-state, cross-word triphones, with 8 Gaussians per state, based on standard MFCC-plus-derivatives. After testing the recognizer on 1 hr of the available data using 3-gram language model, the obtained word accuracy performance was 87%. The employed speech recognition is grammar-based: a grammar instead of an n-gram language model constitutes a robust solution in small tasks, like the examined HRI task in which there is less domain dependent data available for language training. For extra robustness, we first denoise by delay-and-sum beamforming of 8 MEMs channels arranged on a $4\ cm$ linear setup. The beamformer is steered vertically to the platform to enhance the speech signals coming from the front user area.

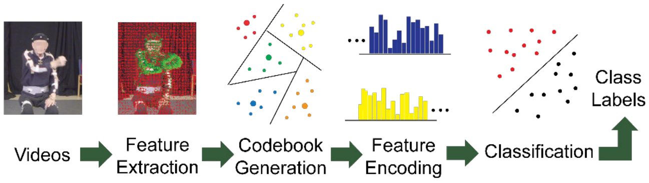

Fig. 4: Action classification pipeline.

Adaptation: The targeted acoustic environment as well as the quality of speech of the elderly users constitute the process of acoustic modeling quite challenging. In the absence of domain specific data, the German triphones are adapted to a development set consisting of recordings from the MOBOT corpus. We use the global maximum likelihood linear regression (MLLR) adaptation technique to transform the means of the Gaussians on the states of the models. The adapted models are then employed for keyword spotting and command recognition as described in our previous work [23].

4. VISUAL GESTURE-ACTIONS

The employed pipeline is depicted in Fig. 4. We employ the terms "action" or "gesture" interchangeably, since the employed method can be applied in both cases or other more complex events.

Feature extraction and descriptors: Dense trajectories [4] 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 [2].

[2] Descriptors include: the Trajectory descriptor, HOG [25], HOF [26] and MBH [4] computed on both axes (MBHx, MBHy). Trajectory descriptor encodes the shape of the trajectories. HOG describes the local static appearance based on the orientation and magnitude of the image intensity gradient. HOF captures motion information using the orientation and magnitude of the optical flow. MBHx/MBHy are computed on the gradient of the horizontal/vertical optical flow components and MBH is their concatenation.

Task	Models	Users									avg.
		p1	p4	p7	p8	p9	p11	p12	p13		
MOBOT-6a	baseline	87.50	60.87	41.67	58.06	66.67	45.45	70.83	81.25		64.03
	+MLLR	96.88	78.26	79.17	90.32	80.00	81.82	79.17	87.50		**84.14**
		u1	u2	u3	u4	u5	u6	u7	u8	u9	
MOBOT-3a	baseline	54.29	33.33	59.05	28.85	36.63	55.34	57.14	52.38	48.08	47.23
	+MLLR	90.48	87.62	96.19	88.46	90.10	96.12	97.14	98.10	90.38	**92.73**

Table 1: Audio classification accuracies for leave-one-out experiments on tasks: (a) MOBOT-6a and (b) MOBOT-3a. The baseline refers to non-matched acoustic models trained on the large vocabulary German corpus task; then the performance is increased by MLLR adaptation.

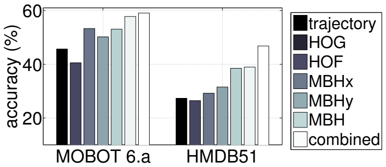

Fig. 5: Classification accuracy on the MOBOT-6a and HMDB51 datasets for multiple descriptors. On MOBOT-6a average accuracy over all patients is shown.

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 [25]. In addition, different descriptors are combined in a multichannel approach, by computing distances between BoVW histograms as:

$$K\left(\mathbf{h}_i, \mathbf{h}_j\right) = \exp\left(-\sum_c \frac{1}{A_c} D\left(\mathbf{h}_i^c, \mathbf{h}_j^c\right)\right), \qquad (1)$$

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.

5. MULTIMODAL FUSION

The late multimodal fusion scheme is as follows: First, the audio and visual scores are normalized to have zero mean and standard deviation equal to one. Then, given the scores for all classes, spoken commands or gesture-actions, the N-best audio gesture commands $g_i, i \in 1 \dots N$ are rescored by combining their scores S_a^i with the corresponding visual ones S_v^i. As a result, the g_i are resorted based on multimodal scores S_{av}^i, obtained by applying a weighted linear combination:

$$S_{av}^i = w_a * S_a^i + w_v * S_v^i \qquad (2)$$

with tunable weights for the audio (w_a) and visual (w_v) modalities. The best multimodal hypothesis g_j is selected as the one with the maximum audio-visual score S_{av}^j where $j = arg \max_i S_{av}^i$. The N parameter corresponds to the number of audio hypotheses employed for rescoring.

6. EXPERIMENTS

Single- and multi-modal gesture classification experiments are carried out on a subset of the MOBOT-6a dataset by including 8 subjects and 8 gestures [3], without limiting the generality of results[4]. Single-modality results on supplementary data are also reported for comparison.

6.1. Audio modality results

We first evaluate the audio based classification on the MOBOT-6a task. As seen in Table 1, the achieved average accuracy on the conducted leave-one-out classification experiments are 84.14%. There is an improvement of 20 percentage points (pp) after adapting the acoustic models to the testing environment.

The current setup is also evaluated on a supplementary audio-only dataset, that is the MOBOT-3a which consists of recordings from 9 adult users operating by voice the MOBOT platform while holding and following the rollator. These results are higher reaching a 92.73% accuracy. As expected, this performance is better due to the fact that the speakers are closer to the microphones and the speakers are not elderly patients as in the MOBOT-6a task, but typical adults; their pronunciation and quality of speech is better and more matched with the training data of the acoustic models. Overall, speech appears to be a dominant modality for the examined tasks and the achieved performance renders the system usable as stand-alone or combined with other modalities. However, as seen in the tougher task, when it comes to the HRI assistive task for the elderly the performance drops, and thus we have a lot to expect by the fusion with the visual modality.

6.2. Visual modality results

MOBOT-6a experimental setup and results: We first extract dense trajectories using the default parameters [5]. For the encoding stage we generate a codebook of $K = 4000$ visual words per descriptor, learned with K-means using 100000 randomly selected training features. Each feature is assigned to its closest visual word. The regularization term of the SVM is $C = 100$.

Table 2 shows the accuracy of our action classification system on each patient for MOBOT-6a. The combined descriptor computed with (1) is employed in these experiments. Results show that the large variability of the gesture performance among patients has a great impact. Figure 5 depicts the mean accuracy over all patients for each descriptor. The combined descriptor performs better, since it encodes complementary information extracted from the RGB channel. In all following experiments we use the combined descriptor.

Supplementary dataset and results: The HMDB51 action dataset [27] lists 6766 videos from movies and YouTube. It is a

[3] The 8 selected gestures are: “Help”, “WantStandUp”, “PerformTask”, “WantSitDown”, “ComeCloser”, “ComeHere”, “LetsGo”, “Park”.

[4] Experiments on in-house gesture data based on the same vocabulary of 19 gestures have shown that our classifier generalizes effectively when gradually increasing the number of gestures from 8 to 19. Particularly, the classification accuracy drops from 93.04% (for 8 gestures) to 84.74% (19 gestures) but still the performance renders the system usable.

[5] Dense trajectories of length L frames are extracted and descriptors are computed within space-time volumes of size $N \times N \times L$ aligned with each trajectory. This volume is subdivided to a spatio-temporal grid of size $n_\sigma \times n_\sigma \times n_\tau$ and descriptors are computed in each grid's cell separately. The final descriptor is formed by concatenating individual cells' descriptors. Here we employ $L = 15$, $N = 32$, $n_\sigma = 2$, $n_\tau = 3$ as in [4].

p1	p4	p7	p8	p9	p11	p12	p13	avg.
40.62	65.22	50.00	64.52	66.67	63.64	70.83	34.38	56.98

Table 2: Visual classification accuracy per patient, MOBOT-6a dataset. "avg." stands for average accuracy over patients.

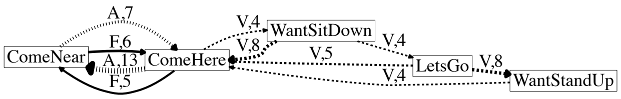

Fig. 6: Multimodal confusability. Nodes correspond to multimodal commands. Archs link a multimodal gesture, e.g. "COME HERE", to the ones confused to, e.g. "COME NEAR" because of an error by one of: audio (A), visual (V), fused (F). Note the number of confusion cases (shown if > 4), to which the linewidths are proportional to. F corresponds to multimodal errors.

REF	ComeHere	WantSit	ComeHere	Park
A	*ComeNear* (0.75)	WantSit (0.58)	*ComeCloser* (0.26)	Park (0.42)
V	ComeHere (0.63)	*ComeHere* (0.45)	*WantSit* (0.35)	Park (0.94)
AV	ComeHere (0.72)	WantSit (0.56)	ComeHere (0.26)	Park (0.44)

Table 3: Examples of correct (in green) multimodal (AV) classification in cases of incorrect (italics in red) audio (A) and/or visual (V) based classification. The best hypothesis is shown for each modality along with their scores. "REF" stands for the ground truth. In the third column, the correct result ("ComeHere") was the 2nd best audio hypothesis, but it was chosen after rearranging the audio hypotheses using the audio-visual scores.

	p1	p4	p7	p8	p9	p11	p12	p13	avg
A	96.87	78.26	79.16	90.32	80.00	81.81	79.16	87.50	84.13
V	40.62	65.21	50.00	64.51	66.66	63.63	70.83	34.37	56.98
AV	87.50	100.0	79.16	96.77	86.66	90.90	95.83	84.37	**90.15**

Table 4: Multimodal (AV) human action classification on the MOBOT-6a data. The N-best list rescoring fusion is compared to the unimodal classification results of the audio (A) or visual (V) streams.

very challending dataset since it has 51 action classes and ranging from facial actions (e.g. smile) to body movements (e.g. stand) and actions are performed in uncontrolled settings. We use the original setup with 3 splits and report average classification accuracy. Figure 5 depicts classification accuracy on HMDB51 and MOBOT-6a. Performance on HMDB51 is consistent with the reported one in [4] using similar settings and is placed within the state-of-the-art (e.g. 55.27% in [28]). Results on both datasets corroborate our choice of the combined descriptor in our experiments.

6.3. Multimodal fusion results

Fusion is evaluated on the MOBOT-6a dataset following the same leave-one-out setup.

Parameters: The fusion parameters are optimized in subsets of the training sets by considering the average performance across the conducted leave-one-out experiments. The optimum values were $(N, w_a, w_v) = (2, 0.95, 0.05)$ after searching the N parameter in the range of values $[2, 8]$ and testing the audio weight w_a using steps of 0.05. Indicative obtained average accuracies in the training sets for $w_a = \{0, 0.6, 0.8, 0.95, 1\}$ are $\{81.3, 83.2, 89.0, 91.2, 85.2\}$. The optimal value $N = 2$ indicates that the correct hypothesis can be found frequently between the 1st or 2nd best audio hypothesis. This is not to underestimate the visual modality as shown next, which is expected to provide supplementary information e.g. when the audio evidence is not enough.

Qualitative results: The complementarity of the audio-visual modalities and the benefits of fusion are justified by the exploration shown in Table 3, showing cases where fusion achieved better. In the first example of the first column, and as mentioned in Fig. 1, although the audio confuses "Come here" with "Come near" which are similar, the error is recovered after the fusion because of the more reliable visual output. Accordingly, in the second column, the gesture "I want to sit down" is confused with "Come here". However the fusion result is correct due to the audio. Finally, fusion may recover errors from both modalities as seen in the third column. Supplementary intuitive confusability results are shown in Fig. 6.

Classification results: As shown in Figure 4, the proposed scheme yields a significant relative improvement of 61% compared to the visual based gesture classification performance (56%). However, the visual modality benefits audio in the fusion process by increasing its performance from 84% to 90%. Overall, the achieved performance in the challenging MOBOT-6a scenario renders the multimodal framework a robust solution.

7. CONCLUSION

In our ongoing work we also focus on the incorporation of the presented approaches in an integrated system on the online processing MOBOT platform employing the robotic operating system (ROS). Based on the multimodal inputs we detect and recognize the issued audio-visual command addressed to the robotic assistant: e.g. the elderly user calls the system by uttering a keyword ("MOBOT") and then provides a voice command along with a gesture. The audio part includes, in addition to the described components, an always-listening one. This is built by: a) voice activity detection, b) keyphrase detection based on the keyword-filler approach, to identify an activation phrase, and c) grammar-based automatic speech recognition. This achieves real-time performance with accuracies close to the reported ones. Regarding gesture recognition, we employ an activity detector, excluding static segments and background movements. Currently, the gesture activity detector and the action classification systems operate at approximately 13 and 3.8 fps respectively on Kinect data.

To conclude with, we described a novel robotic system for multisensor signal processing and presented promising results in multimodal human action recognition, in a newly acquired assistive HRI dataset, focusing on elderly subjects. The qualitative and quantitative multimodal recognition results reach on average 90%. Integration on the robotic platform poses several practical issues which have just been briefly mentioned in this paper. All these highlight this newly formed interdisciplinary research direction which poses challenges on multimodal signal processing, modeling, fusion and on assistive robotics.

Acknowledgement: We thank K. Maninis and G. Pavlakos for their contributions to action/gesture recognition while at NTUA IRAL, and A. Arvanitakis for his help with the ROS platform.

8. REFERENCES

[1] M. Turk, "Multimodal interaction: A review," *Patt. Rec. Letters*, vol. 36, pp. 189–195, 2014.

[2] 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.

[3] R. Poppe, "A survey on vision-based human action recognition," *Image and Vis. Computing*, vol. 28, no. 6, 2010.

[4] 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.

[5] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comp. Vis. and Im. Underst.*, vol. 117, no. 6, 2013.

[6] 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.

[7] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan, "Multimodal feature fusion for robust event detection in web videos," in *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, 2012, pp. 1298–1305.

[8] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann, "Multimedia classification and event detection using double fusion," *Multimedia tools and applications*, vol. 71, no. 1, pp. 333–347, 2014.

[9] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan, "Zero-shot event detection using multi-modal fusion of weakly supervised concepts," in *Proc. Int'l Conf. on Comp. Vis. & Patt. Rec.*, 2014, pp. 2665–2672.

[10] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. Escalante, L. Sigal, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclaroff, "Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary," in *Proc. of Int'l Conf. on Multimedia Interaction*, 2013, pp. 365–368.

[11] J. Wu, J. Cheng, C. Zhao, and H. Lu, "Fusing multi-modal features for gesture recognition," in *Proc. of Int'l Conf. on Multimedia Interaction*, 2013, pp. 453–460.

[12] I. Bayer and S. Thierry, "A multi modal approach to gesture recognition from audio and video data," in *Proc. of Int'l Conf. on Multimedia Interaction*, 2013, pp. 461–466.

[13] K. Nandakumar, K. W. Wan, S. Chan, W. Ng, J. G. Wang, and W. Y. Yau, "A multi-modal gesture recognition system using audio, video, and skeletal joint data," in *Proc. of Int'l Conf. on Multimedia Interaction*, 2013, pp. 475–482.

[14] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos, "Multimodal gesture recognition via multiple hypotheses rescoring," *J. of Machine Learn. Research*, vol. 16, pp. 255–284, 2015.

[15] M. Miki, N. Kitaoka, C. Miyajima, T. Nishino, and K. Takeda, "Improvement of multimodal gesture and speech recognition performance using time intervals between gestures and accompanying speech," *J. on Audio, Speech, and Music Proc.*, vol. 2014, no. 1, pp. 17, 2014.

[16] K. Bousmalis, L. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Proc. Int'l Conf. on Autom. Face & Gest. Rec.*, 2011, pp. 746–752.

[17] V. Ponce-López, S. Escalera, and X. Baró, "Multi-modal social signal analysis for predicting agreement in conversation settings," in *Proc. of Int'l Conf. on Multimedia Interaction*, 2013, pp. 495–502.

[18] Y. C. Song, H. Kautz, J. Allen, M. Swift, Y. Li, J. Luo, and C. Zhang, "A markov logic framework for recognizing complex events from multimodal data," in *Proc. of Int'l Conf. on Multimedia Interaction*, 2013, pp. 141–148.

[19] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Comp. Vis. and Im. Underst.*, vol. 108, no. 1, pp. 116–134, 2007.

[20] B. Burger, I. Ferrané, F. Lerasle, and G. Infantes, "Two-handed gesture recognition and fusion with speech to command a robot," *Autonomous Robots*, vol. 32, no. 2, 2012.

[21] C. Leroux, O. Lebec, M. B. Ghezala, Y. Mezouar, L. Devillers, C. Chastagnol, J.-C. Martin, V. Leynaert, and C. Fattal, "Armen: Assistive robotics to maintain elderly people in natural environment," *IRBM*, vol. 34, no. 2, pp. 101–107, 2013.

[22] R. G. Boboc, A. I. Dumitru, and C. Antonya, "Point-and-command paradigm for interaction with assistive robots," *Int'l J. of Adv. Robotic Sys.*, vol. 12, no. 75, 2015.

[23] 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.

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

[25] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vis. Conf.*, 2009, pp. 124.1–124.11.

[26] 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.

[27] 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.

[28] 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:1405.4506, 2014.