Audio-Visual Robot Command Recognition

Jordi Sanchez-Riera, Xavier Alameda-Pineda, Radu Horaud

Audio-Visual Robot Command Recognition

D-META'12 Grand Challenge

Jordi Sanchez-Riera
INRIA Grenoble Rhône-Alpes
Perception Team
665 Avenue de l'Europe
38334, Montbonnot, France
jordi.sanchez-riera@inria.fr

Xavier Alameda-Pineda
INRIA Grenoble Rhône-Alpes
Perception Team
665 Avenue de l'Europe
38334, Montbonnot, France
xavier.alameda-pineda@inria.fr

Radu Horaud
INRIA Grenoble Rhône-Alpes
Perception Team
665 Avenue de l'Europe
38334, Montbonnot, France
radu.horaud@inria.fr

ABSTRACT

This paper addresses the problem of audio-visual command recognition in the framework of the D-META Grand Challenge[1]. Temporal and non-temporal learning models are trained on visual and auditory descriptors. In order to set a proper baseline, the methods are tested on the "Robot Gestures" scenario of the publicly available RAVEL data set, following the leave-one-out cross-validation strategy. The classification-level audio-visual fusion strategy allows for compensating the errors of the unimodal (audio or vision) classifiers. The obtained results (an average audio-visual recognition rate of almost 80%) encourage us to investigate on how to further develop and improve the methodology described in this paper.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Applications

General Terms

Algorithms, Datasets, Evaluation

Keywords

Audio-Visual Categorization, Multimodal Learning

[1]Datasets for Multimodal Evaluation of Tasks and Annotations: `http://d-meta.inrialpes.fr` Organizers: Xavier Alameda-Pineda (PhD candidate at Perception Team, at INRIA Rhône-Alpes), Kristiina Jokinen (adjunct Professor of language technology at the University of Helsinki) Dirk Heylen (Professor at University of Twente), Roman Bednarik (interaction technology researcher at the University of Eastern Finland) and Michal Hradiš (PhD candidate at Faculty of Information Technology, Brno University of Technology)

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

1. INTRODUCTION

Humans use several sensory modalities to gather information into order to build an internal representation of the real world, e.g, visual, auditory, tactile, or olfactory sensing. Among these modalities, vision and audition are the most suitable to be used by a robot due to the wide availability of associated sensors, namely cameras and microphones. Nevertheless, the data collected by these sensors are often corrupted by occlusions, reverberations, and bad recording conditions such as poor lighting, competing sound sources, background noise, etc. Hence unimodal interpretation leads to ambiguities. It seems natural to investigate how one can obtain a better understanding of a human action if several modalities are used simultaneously. Indeed, the fusion of information coming from different modalities is likely reinforce the true hypothesis. In addition, several challenging questions arise when dealing with different modalities: At which level of abstraction one should fuse the information coming from physical different sensors? How can one build the concept of multimodal features? how can one fuse features representing different aspects (time, position, appearance, spectrum, etc) of the events to be recognized or categorized?

This work is presented in the framework of the D-META ICMI Grand Challenge aiming at setting up the basis for comparison, analysis, and further improvement of multimodal data annotations and multimodal interactive systems. The main goal of D-META is to foster research and development in multimodal communication and to further elaborate algorithms and techniques for building various multimodal applications. Held by two coupled pillars, method benchmarking and annotation evaluation, the D-META challenge envisions a starting point for transparent and publicly available application and annotation evaluation on multimodal data sets. Five tasks were proposed in D-META: **AVRGR**, recognize gestures addressed to the robot by means of vision and the audio, **AVSR**, detect, localize and track multiple speakers using audio-visual information, **CEP**, estimate the level of engagement in a video-mediated communication, **AVCGR**, recognize conversational gestures in first encounter dialogues and **AVFGR**, recognize feedback gestures in first encounter dialogues.

In this paper we set the baseline for the task audio-visual recognition of human commands. The proposed method combines auditory and visual information for recognition using the publicly available RAVEL data set [2], which contains annotated binocular-binaural recordings.

The remainder of the paper is organized as follows. Section 2 describes the previous work done on the topic and the main contributions of the paper. The learning approach as well as the auditory/visual features are described in section 3. In section 4 are explained the different experiments carried out to validate the fusion methods and in section 5 conclusions and directions for future work are briefly presented. Finally, section 6 is devoted to the discussion of some Challenge organizational aspects.

2. RELATED WORK & CONTRIBUTIONS

As mentioned before, a few choices have to be done when combining different sensory modalities for classification. Indeed, the level of abstraction at which the fusion should be done depends on the targeted application. Most of the existing audio-visual categorization methods perform the fusion either at the feature level [5, 9, 10] or at the classification level [7, 14, 17]. The reader is referred to [11] for a study using support vector machines (SVM) that compares feature-level fusion techniques to classification-level fusion techniques.

An example of a feature-level fusion method is [5]. The idea is to track short-term visual features. These features are associated to the auditory signal in order to construct a joint audio-visual codebook. Used to describe a part of a sequence, the codebook feeds a multiple-instance learning (MIL) paradigm, training a semantic concept detector. Another way to perform audiovisual fusion at a feature level is to map both modalities into the same "audio-visual object" space and to perform clustering [6, 1].

The authors in [9] target a speech detection application, and perform the audio-visual fusion at a feature level as well. Principal component analysis (PCA) features are taken from the face images and Mel Frequency Cepstral Coefficients (MFCC) are the auditory features. Both types of features are then projected in a joint subspace using canonical correlation analysis (CCA). A Gaussian mixture model (GMM) is used for classification.

The authors in [10] target general activity recognition. High dimensional features (around 3000) for video and audio are collected and then reduced using the sequential forward floating selection (SFFS) algorithm. The SFFS is a pruning algorithm that selects most relevant features. With this the dimension is reduced (to 40). Finally a kNN algorithm is used as classifier.

Among the classification-level fusion methods we remark [7], in which the authors experiment different combination strategies for object detection. Visual features are based on texture description and entropy-based variable-size patches. Auditory features correspond to the energy of the signal's gammatone filter bank decomposition. Monocular video and monaural audio are used and there is a strong need of uniform visual background.

Object recognition based SVMs is used in [14] where a probabilistic method combining posterior class probability output by each classifier is proposed; Basically this means that each modality is trained separately and then combined. SIFT descriptors are used as visual features and a commercial speech recognizer is used to classify the incoming audio signal.

A different approach from the ones mentioned so far is described [17] which finds sport highlights using a coupled hidden Markov model (CHMM). Several video features are used such as quantization average motion vectors and color. On the auditory side, the authors chose to use MFCC features. Both these features train a CHMM to perform the classification.

In this paper we introduce a method that compares different learning schemes. The first scheme is based on a bag-of-words (BoW) approach while the second one incorporates temporal structure by means of HMMs. Scene-flow [15] and STIPs [8] are used as visual features and MFCC as auditory features. The fusion is done at the classification stage through a modality-weighting scheme, i.e., pooling. Experimental validation is done using a publicly available data set, in which we set the performance baseline.

3. AUDIO-VISUAL CATEGORIZATION

This section is devoted to the proposed audio-visual command learning approach, which performs classification-level fusion. By means of the scene flow from one side and STIPs on the other side, we are able to describe the visual information (see section 3.1). The auditory information is characterized by standard features used in speech recognition (section 3.2). The learning is performed through two different supervised techniques: without any temporal model (section 3.3) and with a temporal model (section 3.4). Finally, the procedure to combine the output of the unimodal classifiers is described (section 3.5).

3.1 The Visual Descriptor

We used two different visual descriptors. The first one was proposed in [15] and it is based on the scene flow, which is the 3D equivalent of the optical flow [16]. The scene flow is represented by the optical flow plus the depth at each image position. Together with the camera calibration, this is equivalent to a vector field of 3D position and associated 3D velocities. This intrinsic representation is potentially less sensitive to changes of texture and illumination than the intensity images. Moreover, the notion of depth allows to focus on the actor, while discarding any activity from the background. We assume that the actor of interest is the person closest to the camera. This is a reasonable assumption, since it holds in most of the human-robot-interaction applications and on movies. The final descriptor consists on the position and disparity relatives to the actor's face plus the optical flow (see Figure 1 for a detailed example).

The second descriptor, STIPs, was proposed in [8]. This descriptor consists on the histogram of gradients concatenated to the histogram of optical flow (HOG-HOF), applied at Harris 3D interest points. Notice that, while the first descriptor uses stereo-vision and has low dimension (five), the second uses monocular vision and is high dimensional (200). Furthermore, while the former has a semi-dense nature, the latter is sparse in the spatiotemporal domain.

3.2 The Speech Descriptor

The auditory stream is represented by the Mel Frequency Cepstral Coefficients (MFCC). Widely used for speech and sound recognition (see [13, 12]), the MFCC are computed following the three steps: (i) perform the short-time Fourier transform (STFT), (ii) map the power spectrum onto the Mel scale and (iii) take the discrete cosine transform of these mapped powers. The are three main parameters associated with MFCC features. First, the frame size defines the length of the STFT (denoted by W). Second, the frame shift (F)

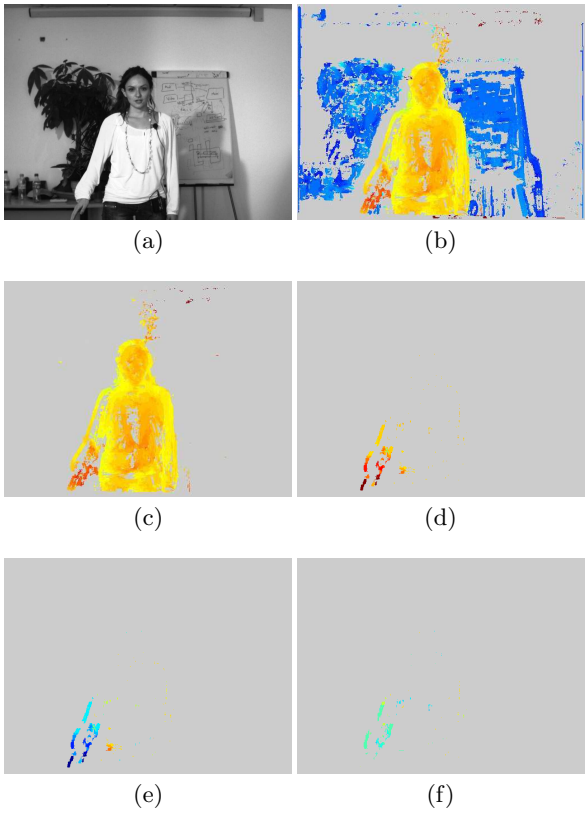


Figure 1: Construction of the proposed descriptor. The actor's face is detected from the left input image (a). The raw disparity map (b) is segmented, such that all pixels having the lower disparity than the actor's face are discarded (c). The descriptor is then computed for all remaining pixels undergoing non-zero motion, such that it consists of the pixel's position relative to the face, it's disparity (d), and horizontal (e) and vertical (f) components of optical flow.

determines the time between two consecutive STFT windows. Third, the amount of cepstral coefficients (D), that sets the dimension of the output MFCC representation.

3.3 Bag-of-Words Gesture Learning

The Bag-of-Words (BoW) paradigm has shown to have very high performance in model learning from low-level features. Following [8], it requires to:

1. Collect a set of local descriptors (possibly associated with interest points) of all training gesture instances.
2. Group these descriptors into K clusters.
3. Quantize all the descriptors assigning the label corresponding to its nearest cluster centroid. This provides for the "words".
4. Represent a gesture instance as a K-bin histogram of the quantized descriptor ("bag of words").
5. Train a classifier with these histograms in order to construct the gesture model.

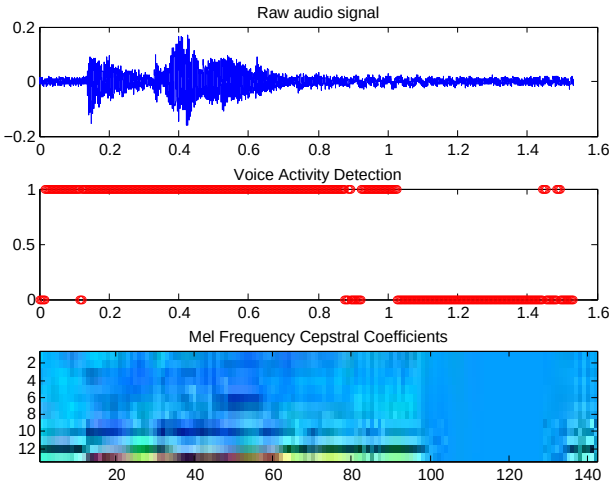


Figure 2: Mel Frequency Cepstral Coefficients for one voice-command instance. From the raw signal (top) the voice activity is detected (middle) and used to mask the extracted MFCC (bottom).

In steps 1–3, the word vocabulary (or the codebook) is constructed. The amount of local descriptors is typically high. Clustering these descriptors allows to represent an instance as a histogram, which in turn provides for a compact representation with fixed (chosen) length K. During steps 4–5, these compact representations together with their annotated labels are used to train a classifier. The BoW representation encodes the relative frequency of occurrences of the quantized descriptors, which discriminates among command classes.

We use the BoW paradigm to build auditory and visual models for each of the commands. Hence we have both a visual and an auditory classifier. When an instance of an unknown command class has to be recognized, the auditory and visual representations are computed and sent to their respective classifiers. The outputs of the two classifiers are fused to perform audio-visual gesture recognition (as explained in section 3.5).

3.4 Temporal Gesture Learning

The standard BoW approach is notoriously lacking a temporal model. Indeed, the order of the extracted features is not taken into account when learning the command model. A powerful statistical learning tool to model time series are hidden Markov models (HMMs). These models assume the existence of a discrete-valued hidden variable that describes the state of the sequence. At each of these states, the model is assumed to generate observations following different probability distributions (called emission distributions).

To fully specify such an HMM, two choices have to be done:

1. The amount of hidden states and the possible transitions among them.
2. The emission probability distributions at each of the states.

An expectation-maximization (EM) algorithm is derived in order to estimate the posterior distributions of the hidden

state sequence, the transition probabilities, and the parameters of the emission distributions (see [12, 3] for more details).

3.5 Audio-Visual Command Recognition

Once the auditory and visual classifiers are trained, we are ready to fuse the information coming from both classifiers. Let $a_c(g)$ and $v_c(g)$ denote the score of the command instance g to belong to class c given by the auditory and visual classifiers respectively. In order to combine the information from both classifiers we train a combined classifier consisting on (i) whitening the training data (unimodal classifier scores) and (ii) apply a weighting function.

The whitening procedure consists on computing the mean (μ_a) and the standard deviation (σ_a) of the auditory classification scores $\{a_c(g_n)\}_{n=1,c=1}^{N,C}$, being N the number of command instances. A new auditory score is computed as $\tilde{a}_c(g) = \frac{a_c(g)-\mu_a}{\sigma_a}$. The same procedure is applied to the visual classification scores.

Finally, the combined score is the result of a convex combination of the two whitened scores:

$$m_c^l(g) = l\tilde{v}_c(g) + (1-l)\tilde{a}_c(g).$$

The value of l determines the trust we put on each modality. Actually, some cases deserve a special mention:

$l = 0$ is equivalent to audio-based classification.

$l = 0.5$ the auditory and visual scores stand on equal foot,

$l = 1$ is equivalent to vision-based classification, and

In general, $l > 0.5$ means that we put more trust on the visual classification score, whereas $l < 0.5$ means that we do it with the auditory score. This way of combining the two classifiers allows us to evaluate the relative trust we put on the modalities. The final classification is:

$$c^* = \arg\max_c m_c^l(g).$$

4. EXPERIMENTAL VALIDATION

In this section we describe the experiments done using different classification methods, such as HMMs and SVMs. The details are given below.

4.1 Implementation

The visual descriptors were computed using for STIPs the default vaules given in the binary code [2] and for Sceneflow using a disparity margin around 5 pixels with the intention to include most of the body of the actor and exclude the rest of the scene (see [15] for other details). The parameters to compute the MFCC features were set to the standard ones in speech recognition: $W = 21.3$ ms, $f = W/2$ and $D = 13$.

We tried the two different learning schemes (non-temporal and temporal) with audio and video features.

sSVM The classifier of the histogram of the scene-flow visual descriptors is a SVM. All the visual descriptors are grouped into $K = 500$ clusters. For each command clip, an histogram of centroid occurrences is computed and normalized. This is used to train the SVM.

lSVM We use the same method as before but on features from [8].

[2]http:// www.di.ens.fr/ laptev/

aSVM The classifier of the histogram of auditory features is a SVM. We trained the SVM using histograms of size $K = 500$ in that case. Since the speech information is sparse in time, a voice activity detector is needed, we chose to use the state-of-the-art voice activity detector available in VOICEBOX [4].

sHMM The per-frame histogram of visual features train temporal models (HMM). Five states per commands and mixture of five Gaussian components were trained. The size of the histogram was $K = 20$.

aHMM A temporal model is added to the auditory features. In that case eight states per command and Gaussian emission probabilities were trained.

lHMM This options was not explored because of the nature of the descriptor. As explained before, the HOG-HOF descriptor is sparse. Actually there are many image frames without any descriptor. Hence this descriptor is not well suited for a per-frame representation.

4.2 The Ravel Data Set

The experimental validation is performed on the "Robot Gestures" scenario of the Ravel data set [2]. We use the eight sequences proposed for the Challenge, each one containing three instances of nine command categories. The set of voice-and-gesture commands are the followings: (i) *wave* ("Hello!"), (ii) *walk towards the robot* ("I am coming."), (iii) *walk away from the robot* ("Bye!"), (iv) *stop* hand-wave ("Stop!"), (v) *turn around* gesture ("Turn around."), (vi) *come here* gesture ("Come here."), (vii) *point* action ("Look!"), (viii) head motion for *yes* ("Yes") and (ix) head motion for *no* ("No"). In all cases, the human gesture is accompanied by speech corresponding to the gesture, shown above in brackets. Notice that all the actors in the data set are non-native English speakers of five different nationalities, hence there is a large variability in the speech commands.

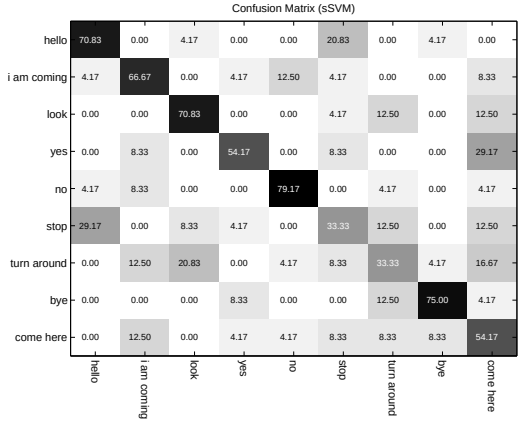
4.3 Performance Results

In order to properly validate the method, the leave-one-out strategy is applied. This strategy is a cross-validation strategy that consists on selecting one of the sequences as test data and perform the training with the rest. Applying this to all the sequences makes the presented results statistically significant.

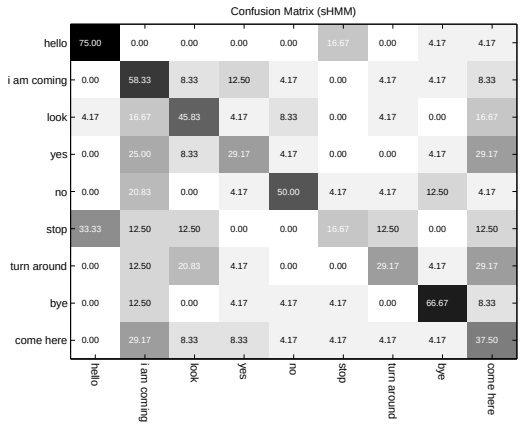
Evaluating multicategory classifiers means providing the confusion matrices. The ij-th entry of such matrix contains how many instances of the i-th class have been classified as class j. By averaging the elements of the diagonal, one obtains the average recognition rate (ARR) of the classifier.

4.3.1 Visual Categorization

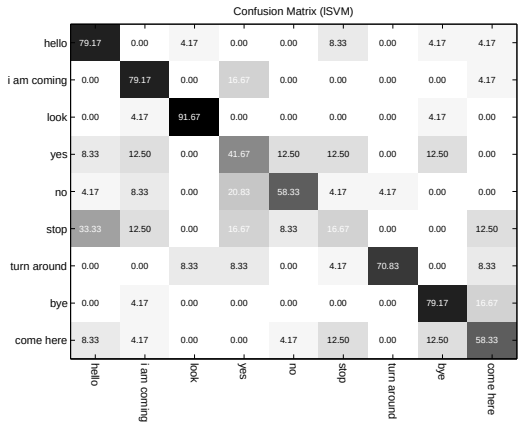
Figures 3a, 3b and 3c show the confusion matrices for the **sSVM**, the **sHMM** and the **lSVM** classifiers. Notice that **sSVM** performs very well in five out of nine gestures (*Hello*, *I am coming*, *Look*, *No* and *Bye*), well in two gestures (*Yes* and *Come here*) and poorly in two gestures (*Stop* and *Turn around*). However, the **sHMM** classifier performs poorly compared to **sSVM**. Indeed, it gets good results for most of the actions, very good results for just two of them (*Hello* and *Bye*) and poor results in three gestures (*Yes*, *Stop* and *Turn*



(a) **sSVM**



(b) **sHMM**



(c) **lSVM**

Figure 3: Confusion matrix of the three visual classifiers: (a) **sSVM**, (b) **sHMM** and (c) **lSVM**.

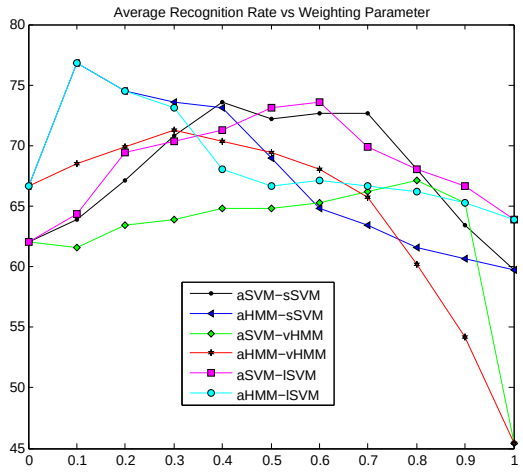


Figure 5: Average Recognition Rate as a function of the multimodal weighting parameter l.

around). Notice also that there is no much difference between the **sSVM** and the **lSVM** classifiers. Actually, they mainly differ in two of the gestures *No* and *Turn around*.

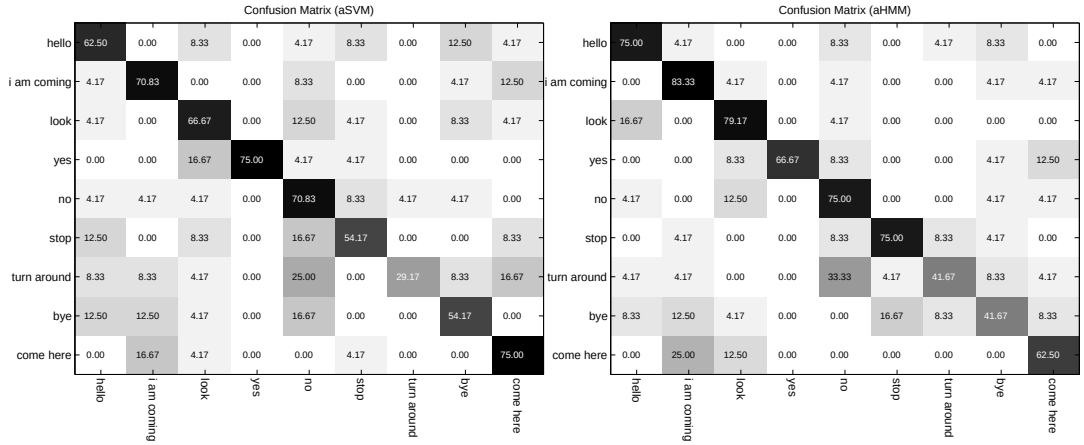
4.3.2 *Auditory Categorization*

The confusion matrix of **aSVM** and **aHMM** can be seen in figures 4a and 4b respectively. The SVM-based classifier performs very good in six out of nine actions, good in two of them (*Bye* and *Stop*) and poorly just in one command (*Turn around*). However the **aHMM** classifier has very good performance everywhere except for the *Bye* and *Turn around* commands.

4.3.3 *Audio-Visual Categorization*

It is worth to notice that, for instance, some actions that are difficult to recognize by **sSVM** as *Stop* or *Come here* are easily classified by **aSVM**, and viceversa with the actions *Bye* or *Hello*. This supports the idea that the combined classifier should outperform both unimodal classifiers. Figure 5 show the ARR of the combined classifier as a function of the weighting parameter l. Please remark that when using the same underlying model (curves **aSVM-sSVM**, **aSVM-lSVM** and **aHMM-sHMM**) the maximum performance of the combined classifier is achieved for values of l around 0.5. However, when the temporal classifier is combined with a non-temporal classifier, the maximum performance of the combined tends to shift towards the modality with temporal modeling. Furthermore, the temporally modeled classifier performs much better when some global information (coming from the non-temporal classifier) is taken into account. In that sense, it is worth to notice that, except for one case (left side of **aSVM-sHMM**), the combined classifiers outperform the unimodal ones when a small weight is given to the other modality's classifier.

Finally, Figures 6a-6f show the confusion matrix of all the multimodal classifiers for the optimal weighting parameter l. Generally speaking, we notice that the performance of these combined classifiers improved with respect to the unimodal ones. Most of them obtain outstanding results for some of the commands and very good results for most of the commands. We need to remark that the **aHMM-sSVM** and the **aHMM-lSVM** classifiers achieve an ARR of 77%,



(a) **aSVM** (b) **aHMM**

Figure 4: Confusion matrix of the two auditory classifiers: (a) **aSVM** and (b) **aHMM**.

that represents a considerable increment respect the ARR computed on each modality independently.

5. CONCLUSIONS & FUTURE WORK

In this work we presented a method to recognize human commands from auditory and visual input signals. Based on a high-performance and solid learning technique, the method uses binocular/monocular visual features and monaural auditory features. Audio-visual recognition is performed at a later stage in which the classification scores from audio and video are combined to get the final audio-visual score, yielding to important increasings on the ARR. The method is tested in leaving-one-out fashion on a publicly available data set. Results show the importance of using both modalities for recognition and leave some room for improving.

More tests can be done trying different type of descriptors for visual and auditory data, as well as, try to fuse these data at different stages.

Acknoledgments

This work was supported by the European project HUMAVIPS FP7-ICT-2009-247525, `http://humavips.eu/`

6. CHALLENGE ORGANIZATION

D-META (Data sets for Multimodal Evaluation of Tasks and Annotations) is the answer that a group of researchers gave to the call for Grand Challenges at the International Conference on Multimodal Interaction (ICMI) 2012. The aim of this challenge is to group researchers working in the same multimodal applications and to encourage them to test their algorithms on a common data set. Obviously, this will set a baseline on the quality of the current methods, the usability of such data sets as well as some hints about the quality of the provided data and annotations.

We would like to remark that the spirit of D-META is far from pursuing a standardization or universal agreement on how to process, present or annotate multimodal data. Instead, we would like to have a deep understanding of the underlying problems related to the use of multimodal data targeting different tasks. Where do the algorithms fail? Why? Which algorithm is best suited for each task?

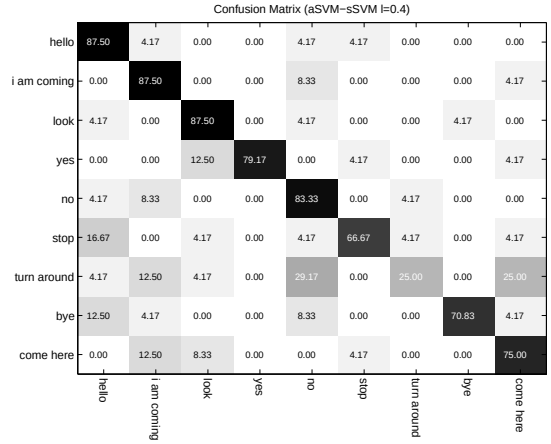
In order to implement those thoughts we selected five different tasks for which we had associated data sets. The tasks were defined by a few researchers working on the field, and they covered a variety of topics: (i) recognition of gestures addressed to the robot by means of the vision and the audio, (ii) detection, localization and tracking of multiple speakers using audio-visual information, (iii) estimate of the level of engagement in a video-mediated communication, (iv) recognition of conversational gestures in first encounter dialogues and (v) recognition of feedback gestures in first encounter dialogues. A call-for-papers was enthusiastically written accordingly to the spirit of the challenge, delineating the proposed tasks and their evaluation procedures, referring to the associated data sets and fitting the quality and logistic criteria of the holding conference, ICMI.

Once the first boost was given and the structure was set, the data set needed to be made public. The data and annotation formats had to be specified too. In addition, the access to this data set had to be free, such that any researcher in the world could participate to the Challenge. Notice that a Golden Standard (the annotation of the Ground Truth) had to be provided for evaluation and transparency purposes.

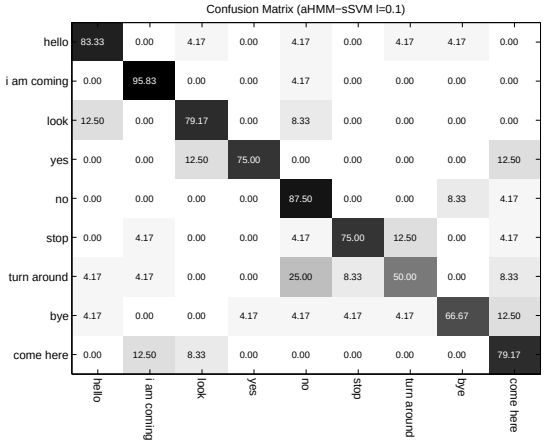
In parallel, a crucial task for the success of such Challenge needed to be done: the advertising. We informed the community about what we were proposing. This is a difficult task because the communication channels are many and not always easy to find and use. Several mailing lists and the ICMI'12 website were used to make the event visible.

We encountered several problems through the organization of this Challenge. First of all, the process of publishing a data set is very resource- and time-consuming. Lots of efforts were made to set a proper access to the data set and their annotations. Secondly, we had also troubles advertising the Challenge. Even if we used all our communication channels, we had almost no external submissions. We think that several factors may have influenced this sad outcome:

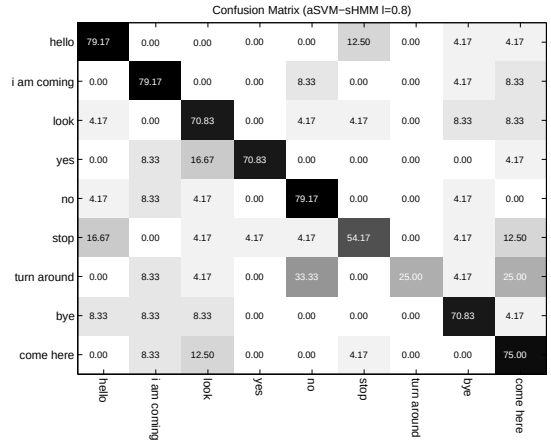
- People are not always willing to test their methods to



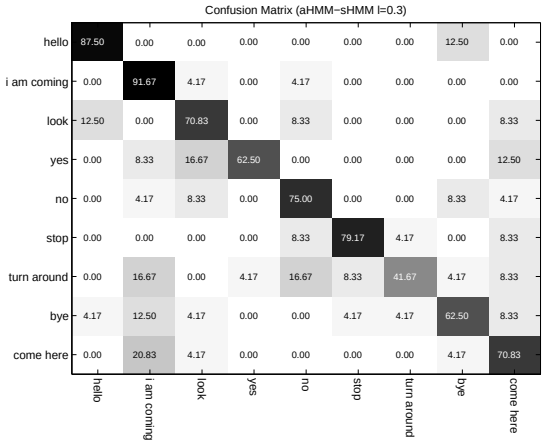
(a) **sSVM-aSVM** ($l = 0.4$)



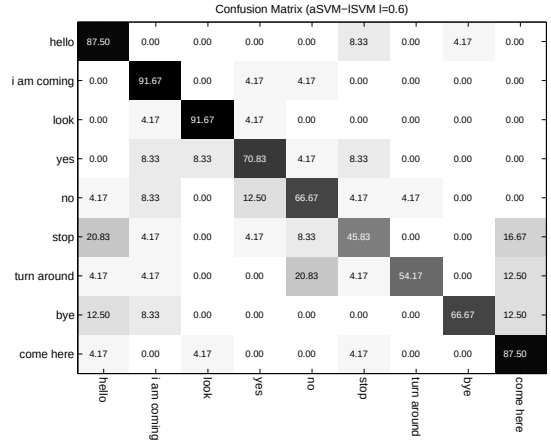
(b) **sSVM-aHMM** ($l = 0.1$)



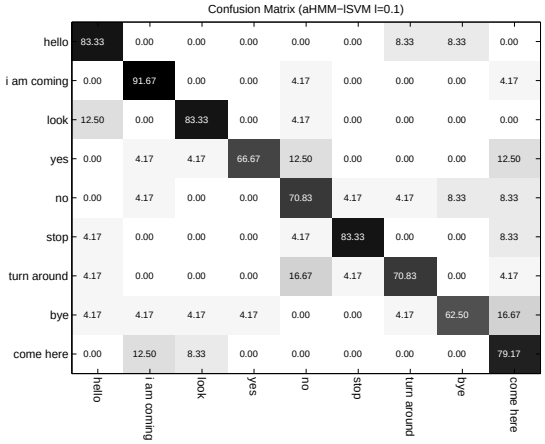
(c) **sHMM-aSVM** ($l = 0.8$)



(d) **sHMM-aHMM** ($l = 0.3$)



(e) **lSVM-aSVM** ($l = 0.6$)



(f) **lSVM-aHMM** ($l = 0.1$)

Figure 6: Confusion matrix of the optimal combined classifiers: (a) **sSVM-aSVM** ($l = 0.4$), (b) **sSVM-aHMM** ($l = 0.1$), (c) **sHMM-aSVM** ($l = 0.8$), (d) **sHMM-aHMM** ($l = 0.3$), (e) **lSVM-aSVM** ($l = 0.6$) and (f) **aHMM,lSVM** ($l = 0.1$).

other data sets/data type without having any guarantees of good performance or a possible award.

- Adapt the software/method to a new data set is not always straightforward, and that may be an obstacle for some researchers.
- Since the Challenge just started (it's its first edition), some people may not find this attractive enough or not even a good way to publish their research, even if we think it is worth.
- We should also consider the possibility that the community is simply not interested in what we propose.

For next editions of the Challenge, we should take into account two important points. On one hand, be aware that the processing and formatting of a data set is a long and tedious task, even for small data sets. The format has to be chosen carefully, since it may constrain the access to the data. On the other hand, take advance on the advertising campaign. The larger the amount of people knows about it, the higher the number of participants and the richer the discussion. We hope to be able to overcome these difficulties for the oncoming editions of D-META.

7. REFERENCES

[1] Xavier Alameda-Pineda, Vasil Khalidov, Radu P. Horaud, and Florence Forbes. Finding audio-visual events in informal social gatherings. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, pages 247–254, Alicante, Spain, November 2011. ACM.

[2] Xavier Alameda-Pineda, Jordi Sanchez-Riera, Vojtech Franc, Johannes Wienke, Jan Čech, Kaustubh Kulkarni, Antoine Deleforge, and Radu P. Horaud. Ravel: An annotated corpus for training robots with audio visual abilities. *Journal of Multimodal User Interfaces*, 2012.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] Mike Brookes. VOICEBOX: Speech processing toolbox for MATLAB. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[5] Wei Jiang, Courtenay Cotton, Shih-Fu Chang, Dan Ellis, and Alexander Loui. Short-term audio-visual atoms for generic video concept classification. In *Proceedings of the 17th ACM International Conference on Multimedia*, 2009.

[6] Vasil Khalidov, Florence Forbes, and Radu P. Horaud. Conjugate mixture models for clustering multimodal data. *Neural Computation*, 23(2):517–557, February 2011.

[7] L. Lacheze, Y. Guo, R. Benosman, B. Gas, and C. Couverture. Audio/video fusion for objects recognition. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[8] I. Laptev. On space-time interest points. *International Journal on Computer Vision*, 64(2-3), 2005.

[9] Ming Liu, Yun Fu, , and Thomas S. Huang. An audio-visual fusion framework with joint dimensionality reduction. In *Proceedings of the IEEE International Conference on Audio Speech and Signal Processing*, 2008.

[10] José Lopes and Sameer Singh. Audio and video feature fusion for activity recognition in unconstrained videos. In *Intelligent Data Engineering and Automated Learning*, 2006.

[11] Jie Luo, Barbara Caputo, Alon Zweig, Jörg-Hendrik Bach, and Jörn Anemüller. Object category detection using audio-visual cues. In *Proceedings of the 6th International Conference on Computer Vision Systems*, 2008.

[12] Lawrence R Rabiner and Ronald W Schafer. *Theory and Applications of Digital Speech Processing*. Pearson, 2011.

[13] V. Ramasubramanian, R. Karthik, S. Thiyagarajan, and Srikanth Cherla. Continuous audio analytics by HMM and viterbi decoding. In *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, pages 2396–2399. IEEE, 2011.

[14] Kate Saenko and Trevor Darrell. Object category recognition using probabilistic fusion of speech and image classifiers. In *Proceedings of the 4th International Conference on Machine Learning for Multimodal Interaction*, 2008.

[15] Jordi Sanchez-Riera, Jan Cech, and Radu Horaud. Action recognition robust to background clutter by using stereo vision. In *In 4th International Workshop on Video Event Categorization, Tagging and Retrieval (VECTaR), in conjunction with IEEE European Conference on Computer Vision*, 2012.

[16] Jan Čech, Jordi Sanchez-Riera, and Radu P. Horaud. Scene flow estimation by growing correspondence seeds. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[17] Ziyou Xiong. Audio-visual sports highlights extraction using coupled hidden markov models. *Pattern Anal. Appl.*, 8(1-2):62–71, 2005.