21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Strategies for multiple feature fusion with Hierarchical HMM: application to activity recognition from wearable audiovisual sensors

Julien Pinquier[1], Svebor Karaman[2], Laetitia Letoupin[2], Patrice Guyot[1], Rémi Mégret[3], Jenny Benois-Pineau[2], Yann Gaëstel[4], Jean-François Dartigues[4]

[1]*IRIT, UMR 5505 CNRS, University of Toulouse, France*
[2]*LABRI, UMR 5800 CNRS, University of Bordeaux, France*
[3]*IMS, UMR 5218 CNRS, University of Bordeaux, France*
[4]*INSERM U.897, University of Bordeaux, France*
{pinquier,guyot}@irit.fr; {karaman,letoupin,benois-p}@labri.fr; remi.megret@ims-bordeaux.fr; {Jean-Francois.Dartigues, Yann.Gaestel}@isped.u-bordeaux2.fr

Abstract

In this paper, we further develop the research on recognition of activities, in videos recorded with wearable cameras, with Hierarchical Hidden Markov Model classifiers. The visual scenes being of a strong complexity in terms of motion and visual content, good performances have been obtained using multiple visual and audio cues. The adequate fusion of features from physically different description spaces remains an open issue not only for this particular task, but in multiple problems of pattern recognition. A study of optimal fusion strategies in the HMM framework is proposed. We design and exploit early, intermediate and late fusions with emitting states in the H-HMM. The results obtained on a corpus recorded by healthy volunteers and patients in a longitudinal dementia study allow choosing optimal fusion strategies as a function of target activity.

1. Introduction

The recognition of activities in video has been mostly dedicated to human actions recognition from external cameras. Activity recognition [1] from the content captured by a wearable camera [4] appears as a new difficult problem that requires the choice of adapted content description and pattern recognition strategies. Indeed, such data includes complex contents, with large motion and unconstrained visual and audio environments. In addition to the traditional analysis of the visual content, audio and motion study has gained attention in the last years. Thus the problematic of fusion from multiple sources has been addressed [9, 10]. The analysis is often performed on pre-segmented data, i.e. shots in edited videos, where the event of interest has only to be recognized. The application on more generic data imposes methods than can simultaneously segment and recognize from the multimedia stream. In particular, HMM offers a classification framework which has been exhaustively studied in various configurations for time based segmentation [1, 2, 4]. Nevertheless, the choice of an optimal description space for the observations remains task dependent and sometimes very empirical. The late advances in concept recognition in complex video content [9, 10] convincingly show that only the fusion of multiple cues allows for increasing the performance of classification/recognition in the case of a strong intra-class variability of scene elements and "events".

Typical multimodality fusion approaches [5, 6] consider a first categorization as feature-level fusion vs decision-level (early vs late) fusion. The former allows capturing specific correlations between modalities, by feature agglomeration and a single training, while the latter facilitates the inclusion of multiple heterogeneous features that are classified separately.

In this paper, we compare these approaches in combination with H-HMM classifiers. We also evaluate a hybrid approach, termed intermediate fusion, which proceed to feature fusion inside the observation model.

The paper is organized as follows: in section 2, the models and fusion strategies are presented. In section 3, the corpus and the experimental evaluation are detailed, and the presented results are discussed.

2. Fusion of multiple features: an HMM framework

2.1. Hidden Markov Model

In [4] we proposed a two-level Hierarchical HMM (HHMM) for recognition of patients' activities in wearable video. Here we recall its main properties. The activities that are meaningful to the medical practitioners, the so-called ADL (Activities of Daily Living) are encoded in the top-level HMM; the set of possible states is defined according to the nomenclature of behavioral patterns. We also introduce a reject state "None" to model non-meaningful observations from doctors' point of view. Thus defined, the top-level HMM (TL-HMM) contains the transitions between "semantic" activities including the reject class. The transitional matrix of the TL-HMM is fixed *a priori* according to the patient's home environment and all initial probabilities are set equal. A bottom-level HMM (BL-HMM) models each activity-state of TL-HMM with *m* non-semantic emitting states, as in [1]. The states of BL-HMMs are modeled by Gaussian Mixture Models (GMM) in the observation space described in the following section. The GMM and the transitions matrix of all the BL-HMMs are learned using the classical Baum Welsh algorithm [2] with labeled data corresponding to each ADL from the nomenclature. For the implementation of the BL- HMMs we use the HTK library [3]. In our experiments, the number of states *m* in BL- HMM is fixed to 3 for ADL states and for the reject class "None". The number of Gaussian Mixture Models was experimentally fixed to 5.

2.2. Multi-modal features

Given the complexity of activities of interest, multiple content description cues have to be explored for their recognition. We therefore design fusion approaches combining visual descriptors and audio descriptors. The descriptors were described in detail in [4] for video, and [7, 8] for audio respectively. To make clear their nature we remind that visual features are the histograms of motion parameters of global camera model, expressing ego-motion of the camera wearer. Furthermore, the cumulated histograms of strong motion changes express the dynamics of activity. Colour features are MPEG7 features expressing mean colour and principle colour contrasts in the visual scene. The localization descriptors and audio descriptors are already conceptual level features, obtained by a preliminary classification. The localization descriptors are obtained from the recognition of the visual models of the rooms of the apartment, represented as probabilities for the patient to be in each room. The audio features are made up of probabilities on speech, music, and silence [7], in addition to probabilities on water flow and vacuum cleaner noises [8]. Each used feature is designed to bring specific and complementary information about the observed activity. The list of descriptors is given in Table 1.

Table 1. Descriptors definition

Dynamic	H_{tpe}	Instant global motion descriptor
	H_c	History of global motion descriptor
	RM	Residual local motion descriptor
Static	CLD	Color Layout Descriptor (MPEG-7)
	Loc	Probabilistic location features (7 classes)
Audio	$Audio$	Probabilistic audio features (7 features)

Hence three description subspaces are designed: the "dynamic" subspace has 34 dimensions, and contains the descriptors $D = (H_{tpe}(x), H_{tpe}(y), H_c, RM)$; the "audio" subspace contains the $k = 7$ audio descriptors $p = (p_1, \ldots, p_k)$; the "static" subspace contains 19 coefficients, more precisely $l = 12$ CLD coefficients $C = (c_1, \ldots, c_l)$ and $m = 7$ localization coefficients $L = (l_1, \ldots, l_m)$. The full description space is thus a Cartesian product $S=D\times p\times C\times L$. In the following we study the strategies of fusion of these descriptors to ensure the best performances of the HHMM classifier.

2.3. Early fusion

We first designed the global description space in an early fusion manner. This refers to the fusion of descriptors preliminary to the classification by concatenating them into an observation vector $o \in \mathbb{R}^N$. Taking into account the relatively low dimensionality of final feature vectors, this fusion is justified even without dimensionality reduction by e.g. PCA method. The dimension of the description space is n = 60 when all descriptors are used. By selecting one or several description subspaces within the 6 defined in Section 2.2 there are (2^6-1) possible configurations to be considered. For our experiments, we used the 13 best configurations, selected from a preliminary evaluation of their performances, (*HtpeAudioLocCLD, HtpeAudioRMLocCLD, HtpeRMCLD, HcLocCLD, HtpeAudioRMCLD, AudioCLD, HcHtpeRMLocCLD, HcHtpeLocCLD, HtpeLocCLD, HcAudioRMLocCLD, HcAudioLocCLD, AudioHcHtpeRMLocCLD* and *RMCLD).* Thus designed the description space is inhomogeneous as it combines features of different types.

2.4. Intermediate fusion

With the intermediate fusion, we will treat the different modalities separately. In this approach, an observation is not the concatenation of all modalities but a set of observations of cardinality equal to the number of modalities. We represent each modality by a "stream", that is a set of measures along the time [3]. More formally, an observation $o_i \in \mathbb{R}^N$ is composed of K modalities. Hence we have K streams of observations $o_{i,1}, \ldots, o_{i,k}$ where $o_{i,k} \in \mathbb{R}^{N_k}$ with $\sum_k N_k = N$. Each state of the HMM models the observations of each stream separately by a Gaussian mixture. Each stream k is weighted by a weight w_{lk} that depends on the activity l which is the same for any state of the same activity with the constraint $\sum_k w_{lk} = 1$. The probability of the observation o_i for the state q_j is:

$$p(o_i, q_j) = \prod_{k=1}^{K} p_k(o_{i,k}, q_j)^{w_{lk}}$$

where $p_k(o_{i,k}, q_j)$ is the probability of the partial observation $o_{i,k}$ of the stream k from the Gaussian mixture associated to stream k of the state j. In our work, each state of the BL-HMM is composed of 3 streams: Audio, Static and Dynamic. In a previous study [11], we obtained equiprobable weights (⅓).

2.5. Late fusion

The late fusion approach uses a HHMM classifier for each subspace of the complete description space and merges the results of these preliminary classifiers for the decision making. In the late fusion scheme in the context of our work a classifier is trained for each modality "Audio", "Dynamic" and "Static". Each modality k is assigned an expertise trust score e_{lk} that is specific to each activity l. This score is defined as a normalized performance computed from the performance measure $perf_{lk}$ obtained by the modality k for the activity l:

$$e_{lk} = \frac{perf_{lk}}{\sum_k perf_{lk}}$$

The distribution of the expertise trust scores, obtained with the F-score performance measure, is presented in the Figure 1. The final decision is computed as follows: for each observation, we allocate a label to an activity by choosing the activity which has the best trusted expert of all modalities.

For each observation o_i, we apply the max fusion operator and thus have:

$$l_i = argmax\left(e_{l_i Audio}, e_{l_i Motion}, e_{l_i Visual}\right).$$

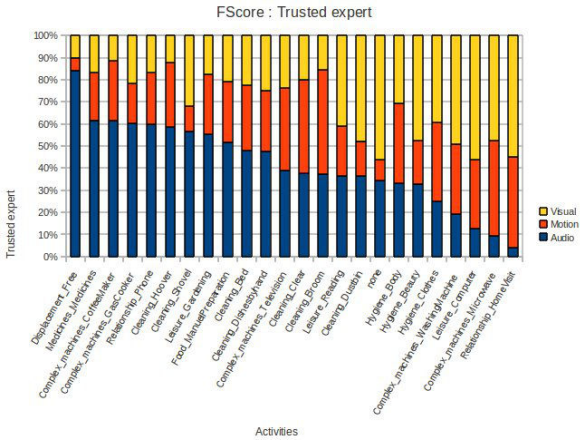

Figure 1. Weights of trusted expert per activity.

3. Experimental evaluation

3.1. Corpus description

The experiments have been conducted on a corpus of videos recorded with our wearable device by patients in their own houses. This corpus contains 37 videos recorded by 34 persons (healthy volunteers and patients) for a total of 14 hours of content. The set of considered activities contains most of the activities already present in the existing methodology for IADL analysis that medical practicians currently use through paper surveys. It is composed of 24 activities in 8 categories, for example "Food: Manual preparation", "Cleaning: Dishes by hand", "Displacement: Free", and "Relationship: Home visit".

3.2. Results

For each fusion strategy, we used leave-one-out cross validation for evaluating the performances. Results are presented in terms of global accuracy, F-score, precision and recall of activities, averaged over the cross validation process. Training was done in a leave-one-out approach at the video level. Training video sequences have been subsampled by a factor of 10 as a preprocessing. The testing has been done on the segments [11] of the last video with the constraint that a segment should contain a minimum of 5 frames and a maximum of 1000 frames. The global results of each fusion approach are presented in Table 2. Theses tables show that the intermediate fusion outperforms the early and the late fusion with a global accuracy of 0.442. Early fusion is almost as good in terms of recall, but lacks precision. These outperform individual descriptor subspaces, which have 0.111 accuracy (Audio and HcHtpeRM) and 0.207 accuracy (LocCLD).

Table 2. Global performances for early, intermediate and three variants of late fusion

Metrics (averaged)	Early fusion	Interm. fusion	Late Fusion		
			F-score trust	Prec. trust	Recall trust
Accuracy	0.207	**0.442**	0.215	0.188	0.210
Precision	0.174	**0.267**	0.106	0.097	0.131
Recall	0.284	**0.288**	0.171	0.155	0.221
F-Score	0.180	**0.253**	0.109	0.092	0.139

A more detailed view of the results is presented in Figure 2, which provides class specific performances in terms of F-score, for each fusion approach. Intermediate fusion has consistently better performances over the various categories with very few exceptions. In this plot, we can observe that some activities such as "Leisure: Computer", "Complex Machines: Hoover", "Complex machines: Gas Cooker" give good results whereas some activities such as "Displacement: Free", "Hygiene: clothes", "Relationship: Home visit" give bad results. These performances seem to reflect that, for the latter activities, the ground truth is not very large. Indeed, this should impede the generalization abilities of the classifiers.

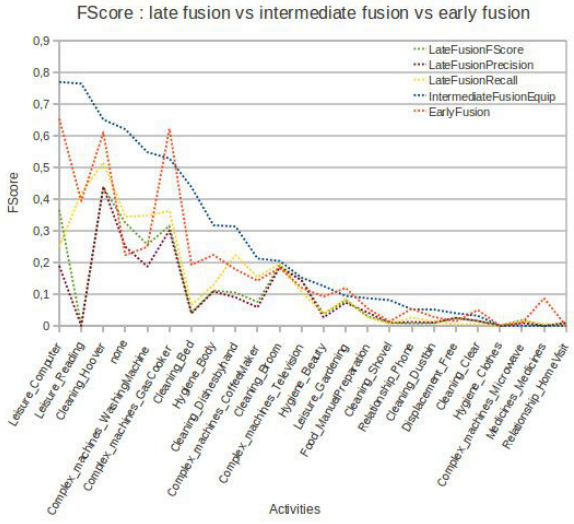

Figure 2. F-score per activity.

4. Conclusion and perspectives

This article has presented a comparison on several approaches for fusion of multiple features. It has been evaluated experimentally on the problem of activity detection in challenging real life audiovisual data, taken from a wearable camera. In the early fusion, we used the concatenation of the description space. In the intermediate fusion, we introduced the concept of a stream of modality: audio, dynamic and static features, extracted from video. In the late fusion, we used the decision scores computed by the Hierarchical HMM for each modality separately. Overall, the experiments have shown that the intermediate fusion has provided consistently better results than the other fusion approaches, on such complex data, supporting its use and expansion in future work.

5. Acknowledgments

This work is partly supported by a grant from the ANR (Agence Nationale de la Recherche) with reference ANR-09-BLAN-0165-02.

6. References

[1] D. Surie, T. Pederson, F. Lagriffoul, L-E. Janlert and D. Sjölie. "Activity Recognition using an Egocentric Perspective of Everyday Objects", Ubiquitous Intelligence and Computing 2007, Springer, pp. 246-257.

[2] L. Rabiner. "A tutorial on hidden markov models and selected applications in speech recognition", Proceedings of the IEEE, volume 77, number 2, pp. 257-286, 1989.

[3] HTK Web-Site: http://htk.eng.cam.ac.uk

[4] S. Karaman, J. Benois-Pineau, R. Mégret, V. Dovgalecs, J. F. Dartigues and Y. Gaestel. "Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases", International Conference on Pattern Recognition, 2010, Istanbul, Turkey.

[5] P. K. Atrey, M. A. Hossain, A. El Saddik and M. S. Kankanhalli. "Multimodal fusion for multimedia analysis: a survey", Multimedia systems, volume 16, number 6, pp. 345-379.

[6] Ayache, S., G. Quénot, and J. Gensel. "Classifier Fusion for SVM-based Multimedia Semantic Indexing", European Conference on IR Research, pp. 494-504, 2007.

[7] J. Pinquier and R. André-Obrecht. "Audio indexing: Primary components retrieval - robust classification in audio documents". Multimedia Tools and Applications (2006), volume 30(3), pp. 313-330.

[8] P. Guyot, J. Pinquier and R. André-Obrecht. "Water flow detection from a wearable device with a new feature, the spectral cover", Content-Based Multimedia Indexing, 2012, Annecy, France, pp. 139-142.

[9] D. Gorisse and al., "IRIM at TRECVID 2010: Semantic Indexing and Instance Search", TRECVID 2010.

[10] Z.-Z. Lan, L. Bao, S.-I Yu, W. Liu and A. G. Hauptmann. "Double Fusion for Multimedia Event Detection", The 18th International Conference on Multimedia Modeling, pp. 173-185, 2012.

[11] Svebor Karaman, Jenny Benois-Pineau, Vladislavs Dovgalecs, Rémi Megret, Julien Pinquier, Régine André-Obrecht, Yann Gaestel, Jean-François Dartigues. "Hierarchical Hidden Markov Model in Detecting Activities of Daily Living in Wearable Videos for Studies of Dementia", Multimedia Tools and Applications, 2012.