This document is published in:

Corchado, E. et al. (eds.) (2010) *Hybrid Artificial Intelligent Systems: 5th International Conference, HAIS 2010, San Sebastián, Spain, June 23-25, 2010. Proceedings, Part II*. (Lecture Notes in Computer Science, 6077). Springer, pp. 436-443. DOI: http://dx.doi.org/10.1007/978-3-642-13803-4_54

Fusion of Single View Soft k-NN Classifiers for Multicamera Human Action Recognition

Rodrigo Cilla, Miguel A. Patricio, Antonio Berlanga, and Jose M. Molina

Computer Science Department, Universidad Carlos III de Madrid
Avda. de la Universidad Carlos III, 22. 28270 Colmenarejo, Madrid. Spain
{rcilla,mpatrici}@inf.uc3m.es, {aberlan,molina}@ia.uc3m.es

Abstract. This paper presents two different classifier fusion algorithms applied in the domain of Human Action Recognition from video. A set of cameras observes a person performing an action from a predefined set. For each camera view a 2D descriptor is computed and a posterior on the performed activity is obtained using a soft classifier. These posteriors are combined using voting and a bayesian network to obtain a single belief measure to use for the final decision on the performed action. Experiments are conducted with different low level frame descriptors on the IXMAS dataset, achieving results comparable to state of the art 3D proposals, but only performing 2D processing.

1 Introduction

Human Action Recognition (HAR) from video is one of the most active research areas in computer vision. Different surveys of the works in the area have been published during the last years [1]. Applications of HAR systems range from video surveillance [2] and Ambient Assisted Living [3] to automatic annotation of video contents [4].

The recognition of Human Actions from video may be considered as a pattern recognition problem [5]. First, a low level descriptor is computed to try to capture the variance on the input frames. Popular choices at this level are motion templates [6], optical flow descriptors [7], spatio-temporal interest points [4], trajectories [8] or a combination of them [9,2]. This computed descriptor is introduced into a classifier to obtain the action category it belongs to. Common choices include Mixtures of Gaussians [8], Support Vector Machines [10], database searches [7,9,2] or Hierarchical Bayesian Models [11]. A particular feature in the recognition of Human Actions is that the actions do not happen isolated, they happen in a temporal sequence. The most popular technique to model the temporal sequence statistics has been Hidden Markov Models [12]. Other proposed techniques have been Context Free Grammars [13] or Conditional Random Fields [14]. In this work we assume that actions happen isolated, focusing on the descriptor classification level.

Most of the existing approaches to HAR have considered a single video sensor to perceive the environment where the actions take place. A single sensor may not be enough to accurately perceive the actions, due to the presence of occlusions.

These occlusions may be caused by the relative position of the human body and the camera (self-occlusions) or by the presence of walls and furniture in the environment. To deal with these problems, HAR systems may be improved using a Visual Sensor Networks (VSN) [15] with overlapped cameras.

In this paper we study how to obtain a single classification of the action perceived by all the cameras from the outputs of a set of single camera soft classifiers. Single camera soft classifiers provide a posterior for the performed activity based on the frame descriptor previously computed. We try two different approaches to solve the problem: the first one is based on a weighted voting scheme; the second one is based on using a Bayesian Network to model the error produced by each one of the single view classifiers. Our approach avoids computing the 3D visual hull, an expensive and centralized task used by state of the art methods for multiple view human action recognition [16,17], using only 2D pattern recognition techniques.

Paper is organized as follows: on section 2 the problem to solve is formally defined; on section 3, the classifier fusion algorithms to be tested are presented; on section 4, we present the single view soft classifier we use to test the classifier fusion algorithms; on section 5, results of applying the proposed algorithms to classify the IXMAS dataset are shown; finally, on section 6, the conclusions of this work are presented.

2 Problem Statement

Let $f_t = \left\{f_t^1, \ldots, f_t^C\right\}$ be a set of action descriptors computed by a set of C cameras at an arbitrary instant t. The posterior probability $p\left(y_n \mid f_t^c\right)$ of action y_n, $y_n \in \mathbf{Y} = \{y_1, \ldots, y_N\}$ is obtained applying a soft classifier to the descriptor f_t^c. Let $B = \{p\left(y_n \mid f_t^c\right)\} \forall n, c$ be the set of all the posterior probabilities obtained after applying the soft classifier to each one of the views. The problem we want to solve is how to combine the single camera posteriors in B into a single posterior for all the cameras, $p\left(y_n \mid f_t^1, \ldots, f_t^1\right)$, $y_n \in \mathbf{Y}$, in order to decide what is the activity y_n being performed.

3 Fusion of Soft Classifiers

Two different algorithms are going to be tested for this task. The first one, a voting scheme. The second one, a bayesian network modeling the errors on local classifications.

3.1 Voting

The first algorithm we are going to test for the fusion of single view soft classifications is defined to be the sum of the posterior probabilities.

$$p\left(a_i \mid f_t^1, \ldots, f_t^C\right) \propto \sum_{c=1}^{C} p\left(a_i \mid f_t^c\right) \tag{1}$$

3.2 Bayesian Network

The second algorithm we are going to test for the fusion of single view soft classifications is based on the Bayesian Network shown on figure 1. The network is composed of observation nodes f_t^c, representing the observation at instant t on camera c, a node α_t representing the activity at time t and a set of latent nodes v_t^c, to model the single view classification.

Given a set of frame descriptors $\mathbf{f}_t = f_t^1, \ldots, f_t^C$, a set of latent variables $\mathbf{v}_t = v_t^1, \ldots, v_t^c$, and the activity label α_t, their joint probability is factorized as:

$$p\left(\alpha_t, \mathbf{v}_t, \mathbf{f}_t\right) = p\left(\alpha_t \mid \mathbf{v_t}\right) \prod_{c=1}^{C} p\left(v_t^c\right) p\left(f_t^c \mid v_t^c\right) \tag{2}$$

The probability of α_t is defined as a product of independent factors, assuming independence between hidden variables v_t^c:

$$p\left(\alpha_t \mid \mathbf{v_t}\right) \doteq \prod_{c=1}^{C} p\left(\alpha_t \mid v_t^c\right) \tag{3}$$

With this assumption we refuse to model correlations between local classification errors. In this way, when adding a new camera to the system only 2 conditional probability distributions need to be estimated, instead of the exponential number of them if the assumption were not made. Thus, equation 2 can be rewritten as:

$$p\left(\alpha_t, \mathbf{v}_t, \mathbf{f}_t\right) = \prod_{c=1}^{C} p\left(\alpha_t \mid v_t^c\right) p\left(v_t^c\right) p\left(f_t^c \mid v_t^c\right) \tag{4}$$

The posterior probability of an activity label α_t and a set of hidden variables $\mathbf{v}_t$ is proportional to the joint probability:

$$p\left(\alpha_t, \mathbf{v}_t \mid \mathbf{f}_t\right) \propto p\left(\alpha_t, \mathbf{v}_t, \mathbf{f}_t\right) \tag{5}$$

Given a set of frame descriptors $\mathbf{f}_t$, the posterior probability of the activity label α_t is obtained marginalizing equation 5 over the set of latent variables $\mathbf{v}_t$:

$$p\left(\alpha_t = a_i \mid \mathbf{f}_t^c\right) = \sum_{j=1}^{N} \prod_{c=1}^{C} p\left(\alpha_t = a_i \mid v_t^c = a_j\right) p\left(v_t^c = a_j\right) p\left(f_t^c \mid v_t^c = a_j\right) \tag{6}$$

$p\left(f_t^c \mid v_t^c = a_j\right)$ may be computed in terms of $p\left(v_t^c = a_j \mid f_t^c\right)$ using Bayes theorem:

$$p\left(f_t^c \mid v_t^c = a_j\right) = \frac{p\left(v_t^c = a_j \mid f_t^c\right) p\left(f_t^c\right)}{p\left(v_t^c = a_j\right)} \doteq \frac{p\left(v_t^c = a_j \mid f_t^c\right)}{p\left(v_t^c = a_j\right)} \tag{7}$$

The term $p\left(f_t^c\right)$ vanishes assuming that $f_t^c \sim Uniform$. The final expression for the posterior is obtained introducing the RHS of equation 7 into equation 6:

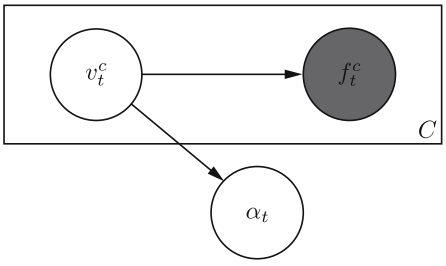


Fig. 1. Plate model of the Bayesian Network used to combine the outputs from the classifiers at each camera

$$p\left(\alpha_t = a_i \mid \mathbf{f}_t^c\right) = \sum_{j=1}^{N} \prod_{c=1}^{C} p\left(\alpha_t = a_i \mid v_t^c = a_j\right) p\left(v_t^c = a_j \mid f_t^c\right) \tag{8}$$

Network parameters are estimated using labeled training samples. $p\left(v_t^c \mid f_t^c\right)$ is known, being provided by the single view soft classifiers, so only $p\left(\alpha^t \mid v_c^t\right)$ needs to be estimated. Be $O^c = (o_1^c, \ldots, o_K^c\}$ the set of K training frame descriptors computed at camera c with their corresponding activity labels $Y^c = \{y_1^c, \ldots, y_K^c\}$, $y_k^c \in A$. Model parameters are estimated as:

$$p\left(\alpha^t = a_i \mid v_t^c = a_j\right) = \frac{\sum_{k=1}^{K} \gamma_k p\left(v_t^c = a_j \mid o_k^c\right)}{\sum_{l=1}^{N} \sum_{k=1}^{K} \gamma_k p\left(v_t^c = a_l \mid o_k^c\right)} \tag{9}$$

where $\gamma_k = 1$ if $y_k = a_j$ and $\gamma_k = 0$ otherwise.

4 Soft Classifier

The classifier we are going to use to obtain the probability of each single frame being an instance of each action category is based on a k-Nearest Neighbor setting (kNN). Let $D = \{x^i, y^i\}$, $1 \leq i \leq M$ be a set of M training samples, being $y_i \in \{y_1, y_N\}$ the label corresponding to the instance x_i. The posterior probability $p\left(y \mid x^j\right)$ of a new sample x_j to predict is decided sampling from the neighborhood of x^i, transforming the distances to the k nearest neighbors into likelihood values:

$$p\left(y = y_n \mid x^j\right) \propto \sum_{k=1}^{K} \gamma_k \left(\rho_j - \left|\left|x^j - x^k\right|\right|\right) \tag{10}$$

where $\rho_j = \sum_{k=1}^{K} \left|\left|x^j - x^k\right|\right|$,i.e. the sum of the distances to the k nearest neighbors of x_j; $\gamma_k = 1$ if $y_n = y_k$ and $\gamma_k = 0$. The main advantage of this classifier is

that it captures the local structure of the data, being able to model multimodal distributions. Training is also very fast because only requires storing the samples on the database.

5 Experiments

5.1 Experimental Setup

Experiments are going to be conducted on the state-of-the art testbed for human action recognition: the Inria IXMAS dataset [1]. The dataset includes samples of eleven action categories performed by 12 different actors 3 times each (36 clips), recorded by 5 different camera views. The actions are: *check watch*, *cross arms*, *scratch head*, *sit down*, *get up*, *turn around*, *walk*, *wave*, *punch*, *kick* and *pick up*. Two different frame descriptors are used to model these actions and test our algorithms. The first one is the popular Motion History Image (MHI) [6]. This descriptor is based on a temporal accumulation of the human body shape. The computed descriptors are resized to a box of 35x20 pixels, obtaining a feature vector of length $l_{MHI} = 700$. The second one is the proposed by Tran et al.[9], including both shape and optical flow information. The extracted descriptor can be obtained from their web[2], being its length $l_{Tran} = 286$.

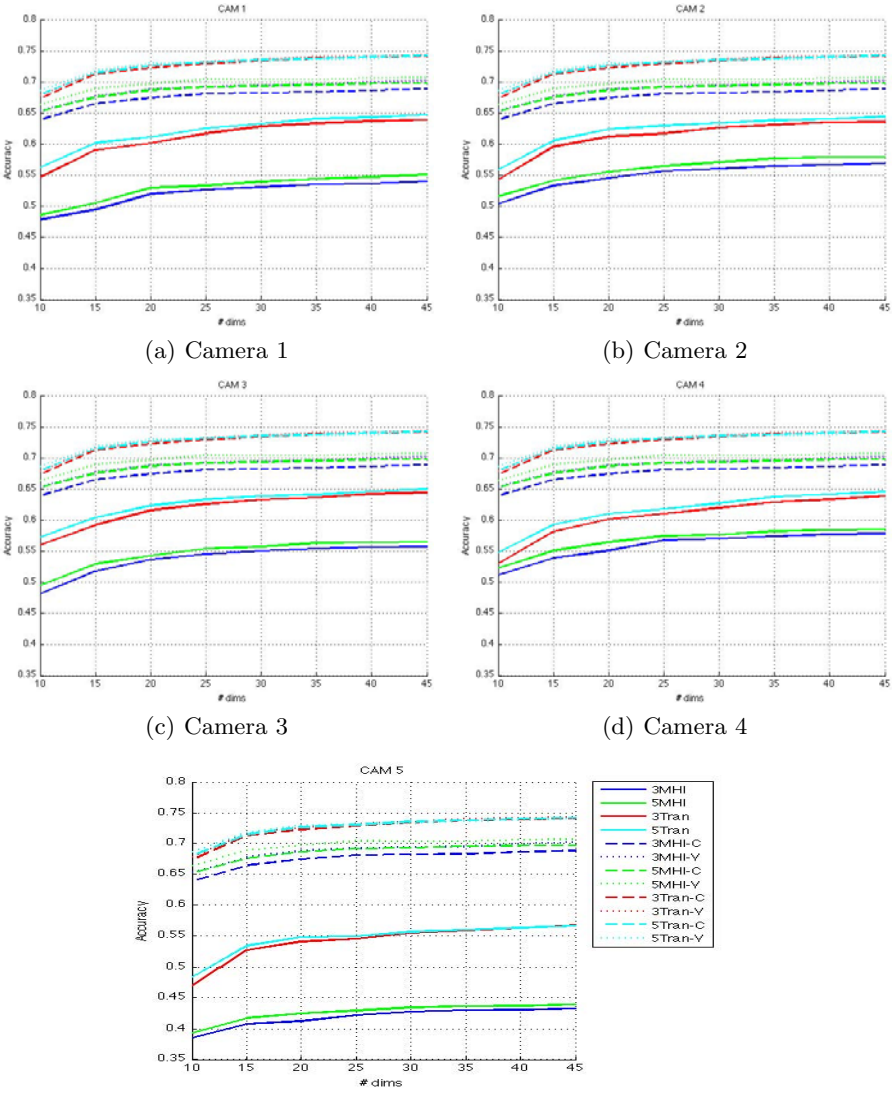
The evaluation protocol to test the classification and fusion algorithms is Leave-One-Clip-Out-Cross Validation: algorithms are trained with all the action clips unless one, that is used for testing. The procedure is repeated until all the clips have been used for testing. The kNN classification algorithms are going to be tested using neighborhood values of $k = 3$ and $k = 5$. As the length of the descriptors is too large for practical usage, the well known Principal Component Analysis is applied to obtain reduced descriptors ranging from $l = 10$ to $l = 45$ with a stepsize of 5.

5.2 Results and Discussion

Single camera classification. Figure 2 shows the result of classifying each one of the camera views with the soft kNN classifier. It is clear that the Tran descriptor predicts the activity performed on a single frame better than the MHI. This behavior was expectable until some point, because Tran's descriptor includes shape and local motion information, while the MHI only includes shape. The classifiers with $k = 5$ always work better than those with $K = 3$.

Single camera classification results also show that while from cameras 1-4 the obtained accuracy is similar, camera 5 accuracy drops about a 10%. Camera 5 provides a top view of the action, preventing descriptors from accurately capturing the dynamics of the performed action.

[1] http://charibdis.inrialpes.fr/
[2] http://vision.cs.uiuc.edu/projects/activity/



(a) Camera 1

(b) Camera 2

(c) Camera 3

(d) Camera 4

(e) Camera 5

Fig. 2. Classification results at each camera before and after the fusion. The first number stands for the number of nearest neighbors used. The suffix stands for the fusion algorithm used: V for voting and C for the bayesian network.

Fusion results. The different plots shown on figure 2 also show the results obtained after applying the fusion algorithms to the single camera soft classifications. The weighted voting proposed on section 3.1 and the bayesian network proposed on section 3.2 have similar results, being voting slightly better. Fusion algorithms improve more the classification based on MHI descriptors. This is

Table 1. Comparison of the accuracy of our method to others

Method	Accuracy	Type
Tran et al. [9]	81	2D
Srivastava et al. [18]	81.4	2D Multicamera
Our	92.01	Multicamera
Weinland et al. [16]	93.33	3D
Peng et al. [17]	94.59	3D

probably because as the initial result was worse than when using Tran descriptors, it is easier to improve the results using fusion.

Comparison to other proposals. Finally, on table 1, we compare the results obtained by our method to the obtained by other state of the art approaches. Our method performs better than other 2D multicamera approaches, obtaining results comparable to proposals based on computing the 3D visual hull. Results on the table are for sequence classification. To obtain them, each frame on a sequence has voted with its posterior distribution to obtain the majority classification.

6 Conclusions

In this paper we have shown how the accuracy of the task of human action classification can be improved combining the results of single view classifiers. We want to remark that our method avoids visual hull computation, being very easy to implement on a distributed environment. Another advantage of the proposed method is that it can integrate other sensors without very much effort, because the fusion level is independent of the type of sensor used. If a posterior for the activity can be obtained from the hypothetical sensor, it can be used in our system.

Future works will explore how to model the correlations between the soft classifications from each camera. We suspect that the independence assumption made between sensor values is too strong, and that fusion results may be highly improved introducing dependencies between sensors in our fusion model.

Acknowledgment. This work was supported in part by Projects CICYT TIN2008-06742-C02-02/TSI, CICYT TEC2008-06732-C02-02/TEC, CAM CONTEXTS (S2009/TIC-1485) and DPS2008-07029-C02-02.

References

1. Lavee, G., Rivlin, E., Rudzsky, M.: Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video. IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews 39(5), 489–504 (2009)

2. Robertson, N., Reid, I.: A general method for human activity recognition in video. Computer Vision and Image Understanding 104(2-3), 232–248 (2006)
3. Cilla, R., Patricio, M., Belanga, A., Molina, J.: Non-supervised discovering of user activities in visual sensor networks for ambient intelligence applications. In: 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies, ISABEL 2009, November 2009, pp. 1–6 (2009)
4. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Miami, USA (2009)
5. Bishop, C., et al.: Pattern recognition and machine learning. Springer, New York (2006)
6. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 257–267 (2001)
7. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision, vol. 2, pp. 726–733 (2003)
8. Ribeiro, P., Santos-Victor, J.: Human activity recognition from video: modeling, feature selection and classification architecture. In: International Workshop on Human Activity Recognition and Modeling, HAREM (2005)
9. Tran, D., Sorokin, A., Forsyth, D.: Human activity recognition with metric learning. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 548–561. Springer, Heidelberg (2008)
10. Cao, D., Masoud, O., Boley, D., Papanikolopoulos, N.: Human motion recognition using support vector machines. Computer Vision and Image Understanding 113(10), 1064–1075 (2009)
11. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 79(3), 299–318 (2008)
12. Oliver, N., Rosario, B., Pentland, A.: A Bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 831–843 (2000)
13. Guerra-Filho, G., Aloimonos, Y.: A Language for Human Action. Computer 40(5), 42–51 (2007)
14. Quattoni, A., Wang, S., Morency, L., Collins, M., Darrell, T.: Hidden-state conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2007)
15. Cucchiara, R., Prati, A., Vezzani, R.: Making the home safer and more secure through visual surveillance. In: Symposium on Automatic detection of abnormal human behaviour using video processing of Measuring Behaviour, Wageningen, The Netherlands (2005)
16. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding 104(2-3), 249–257 (2006)
17. Peng, B., Qian, G., Rajko, S.: View-Invariant Full-Body Gesture Recognition via Multilinear Analysis of Voxel Data. In: Third ACM/IEEE Conference on Distributed Smart Cameras (September 2009)
18. Srivastava, C., Iwaki, H., Park, J., Kak, A.C.: Distributed and Lightweight Multi-Camera Human Activity Classification. In: Third ACM/IEEE Conference on Distributed Smart Cameras (September 2009)