This document is published in:

Corchado, E. et al. (eds.) (2011) *Hybrid Artificial Intelligent Systems: 6th International Conference, HAIS 2011, Wroclaw, Poland, May 23-25, 2011, Proceedings, Part II.*
(Lecture Notes in Computer Science, 6679). Springer, pp. 136-143. DOI: http://dx.doi.org/10.1007/978-3-642-21222-2_17

Improving the Accuracy of Action Classification Using View-Dependent Context Information

Rodrigo Cilla, Miguel A. Patricio, Antonio Berlanga, and Jos 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,{berlanga,molina}@ia.uc3m.es

Abstract. This paper presents a human action recognition system that decomposes the task in two subtasks. First, a view-independent classifier, shared between the multiple views to analyze, is applied to obtain an initial guess of the posterior distribution of the performed action. Then, this posterior distribution is combined with view based knowledge to improve the action classification. This allows to reuse the view-independent component when a new view has to be analyzed, needing to only specify the view dependent knowledge. An example of the application of the system into an smart home domain is discussed.

1 Introduction

The recognition of human actions from video has been a very active research field during the last two decades. Video Surveillance, Multimedia indexing and retrieval or Ambient Assisted Living are some of the applications that have been benefited from the advances made during this period [12].

One of the most recent research issues on human action recognition is the usage of viewpoint independent action representations. The objective is to obtain a representation of the action robust to changes in the viewpoint of the camera that grabs the processed images. The problem has been studied from different perspectives. Some approaches [13,10,6,16] rely on imposing geometrical constraints on 3D body limbs configurations. Its use is limited to situations where an accurate body part tracker is available. Others try to study 3D visual hull evolutions of the human body [22,11,20]. Some proposals associate 2D silhouettes extracted from a single camera with their corresponding 3D visual hulls [8,21]. View-independent action classification can be achieved with non-linear classifiers, [9,19], achieving a promising accuracy. Recently, view-invariant actions representations have been learned in low dimensional manifolds [14,7].

Obtaining a true view invariant representation of the action would allow to share action models between different scenarios, reducing the cost of creating new action recognition systems. The view-invariant action primitives then would be used from a library, only requiring to define scenario dependent action semantics.

By the other hand, the likelihood observing a given action is very dependent on the features of the scene being analyzed. The action *sit* is going to be more

likely to happen in a place where there is a chair, or the action *walk* in a clear area of the scene. The scene context information can help us to reason about what actions are more likely to be performed by the observed human. Combining view-invariant action classifiers with context information might then be a way to improve the final system accuracy.

Context information has been used to improve the performance of different computer vision tasks. Gomez-Romero et al. [5] propose the usage of context information to improve the performance of object tracking systems. Robertson and Reid [15] proposed a probabilistic discriminative framework to combine different attributes of human action recognition, improving optical flow action classification using position and speed knowledge. In a related work, Wu and Aghajan [23] use the actions of the user as the context information to discover the objects in the environment.

This paper presents an action recognition system combining view-invariant classifiers with context information. A viewpoint independent classifier (VIC) is trained using samples of actions obtained from different cameras. For a given view, we learn a context probabilistic model (CPM) relating the positions of the human bounding boxes in the view and the actions. To decide what is the action happening on a new observation the output of the VIC is averaged by the CPM, making the final result.

Paper is organized as follows. In section 2 we present the general structure of the system; in section 3 the view-independent classifier is introduced; in section 4 the view-dependent model used as context is presented; in section 5 we show how the system can be used in a smart home. Finally, in section 6, the conclusions and future research lines of this work are presented.

2 System Overview

The architecture of the proposed system can be observed on figure 1. At a given instant t, a image $I_t(x, y)$ of the scene is grabbed from a fixed viewpoint. A foreground mask $F_t(x, y)$ containing the humans of interest is extracted using background subtraction [17]. The bounding boxes of the objects of interest, i.e., the people in the view, are tracked over time, with a method such the proposed in [4]. For simplicity we restrict to the case of just one person being observed. His bounding box is represented by a 4d vector $b_t = (x, y, w, h)$, where the first two components correspond to the centroid and the last two to the size. The feature extractor (FE) extracts the attribute vector f_t and feds it into the view invariant classifier (VIC) to obtain a preliminary posterior distribution of the action α_t performed, $p(\alpha_t \mid f_t)$, $\alpha_t \in A = \{a_1, \ldots, a_N\}$. This distribution is averaged by the context probabilistic model (CPM) to take into account local information, using b_t. A final posterior probability distribution $P(\alpha \mid b_t)$ is produced to decide on the action most likely to be happening. The system here described does not consider the temporal extents of the actions and restricts to the current instant t for simplicity. The architecture of the system allows sharing the view independent action models across different views, being able to load them from a common

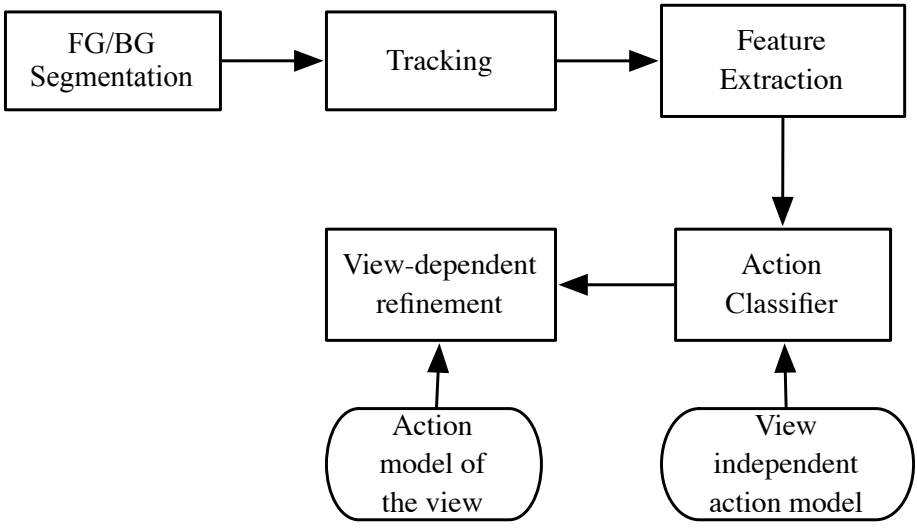


Fig. 1. Overview of the proposed system

library. The system is designed to work with 2D view-independent features, as only the images grabbed from a single camera are used. Also, instead of locating the user in the scene 3D world, it is only located in the 2D view. A similar system could be designed to process different simultaneous views of the scene and compute 3D view-invariant features, such the derived from visual hulls. The location of the user can then be made in the 3D scene world. Note it is not necessary to have multiple views to obtain the 3D location, but we restrict the proposed system to 2D. Including the size of the person bounding box into the context probability distribution might be a naive form of modelling the depth in the scene, as the real size of the person remains constant but not the size of the respective bounding box.

The usage of posterior probability distributions to model the uncertainties makes possible the use of context information. If the view-independent action classifier would produce a crisp output, there would not be any possibility about make an action decision when the output does not agree with the context model. At the same time, a probability distribution can easily relate actions with the places where they are most likely to happen.

3 View-Independent Action Recognition

Producing a view-independent action recognition system is a hard task that is still an open research issue. Some of the efforts already made to accomplish the task were presented on section 1. Before introducing the proposed classifier, it is worth mentioning that it is not a really view-independent classifier. It does

not make any explicit generalization on the viewpoint, just trying to accurately predict the action class of a set of training samples taken from different viewpoints. However, it is enough to show how the performance of the recognition can improve when context knowledge is added.

3.1 Feature Extraction

The combination of optical flow and appearance information has shown to be effective for the recognition of human actions [19]. Optical flow provides information about the current dynamics of the person, while appearance provides information about the current pose. The main problem when computing features for action recognition using the output of an object tracker is the noise, producing random perturbations on the position and size of the bounding box that can generate large variations on feature values.

A good choice in this case can be the usage of histogram feature representations, as they seem robust to perturbations of the bounding box properties. The Histogram of Oriented Optical Flow (HOOF) [2] provides a scale and direction invariant representation of the motion of a target. The Histogram of Oriented Gradients (HOG) [3] is a widely used feature for object detection, providing representation of the object appearance We use the concatenation of both features f_t as the input to the action classifier.

3.2 Action Classification

At was previously mentioned on section 2, the chosen classifier should provide a probabilistic output in order to be able to average the output with the context information. The Relevance Vector Machine (RVM) [18] is a probabilistic model for classification and regression, belonging to the class of sparse kernel methods. RVM automatically selects the basis functions to be used, being very fast to operate because an input pattern need to be only compared with a few samples stored during training. The details on the derivation and implementations of this model are out of the scope of this paper. Readers are encouraged to check [18] for more details. For us, the RVM is a black box that once trained provides an estimation of the probability $p(\alpha_t \mid f_t)$ of an action α_t given the observed feature f_t.

To use the RVM, we need to specify the kernel function to be used to compare two samples x and y. The form of the descriptor introduced in the previous section is an histogram. This motivates us to choose the χ^2 kernel function:

$$K(x,y) = exp\left(-\frac{1}{2}\sum_{b=1}^{B}\frac{(x^b - y^b)^2}{(x^b + y^b)}\right). \quad (1)$$

where x^b and y^b respectively denote the b bin of two histograms x and y to be compared.

4 Probabilistic Context Model

The PCM is defined to model the existing relationship between the actions and the different zones of the view where they are more likely to be observed. A likelihood is associated to each one of the zones of the image to reflect the plausibility of the action happening there using a generative distribution $p(b_t \mid \alpha_t)$. As the value of the function increases, it would be more likely to observe the action α_t at b_t.

There are different choices to model the generative distribution $p(b_t \mid \alpha_t)$. The most simple is to define a gaussian distribution for each action class, but that can be too rigid in the sense that the actions can be observed only on the neighbourhood of a given zone. That is why a gaussian mixture model seems to be a more appropriate choice:

$$p(b_t \mid \alpha_t = a_i) = \sum_{i=1}^{K_c} \pi_i N(\mu_i, \Sigma_i). \tag{2}$$

where π_i corresponds to the weight of the ith component of the mixture, and $N(\mu_i, \Sigma_i)$ denotes the standard gaussian distribution with mean μ_i and covariance matrix Σ_i. The number of K_c mixture components of each class would be empirically determined during training. The parameters of each one of the gaussian mixtures would be estimated from training samples using the standard Expectation-Maximization algorithm [1].

Other possibility is to use a Kernel Density Estimator [1]. This prevents having to choose a priori the number of components in the mixture. Given a set of training samples $X\left(b^1 \ldots b^{N_i}\right)$ of the ith class, the probability density function for the class is defined as:

$$p(b_t \mid \alpha_t = a_i) = \frac{1}{N_i} \sum_{i=1}^{N_i} K\left(b^i - b_t\right). \tag{3}$$

where $K(.)$ can be any kernel function such a gaussian.

The posterior probability distribution on the action provided by the VIC, $p(\alpha_t \mid f_t)$, is combined with the estimation of the probability of each class on the current area $p(\alpha_t \mid b_t)$, to give a final posterior distribution $p(\alpha_t \mid b_t, f_t)$ of the form:

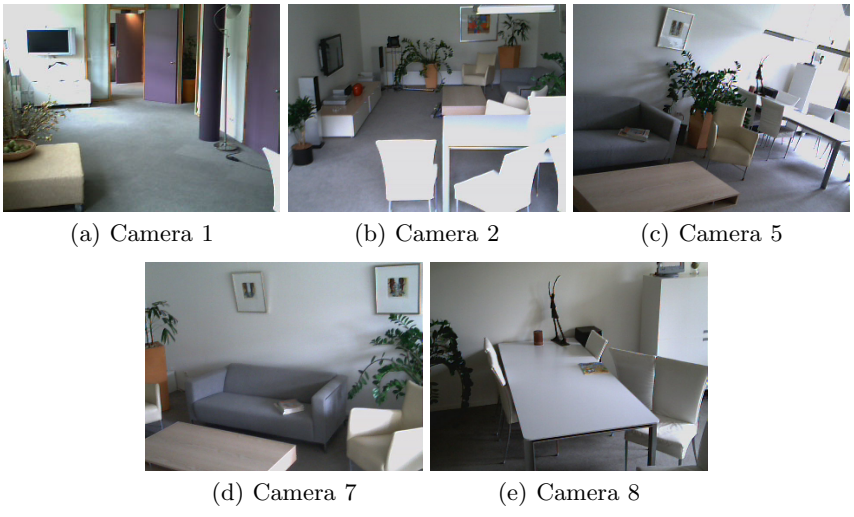
$$p(\alpha_t \mid b_t, f_t) \propto p(\alpha_t \mid b_t)\, p(\alpha_t \mid f_T). \tag{4}$$

It has been assumed that the area b_t and the feature descriptor f_t are independent. $p(\alpha_t \mid b_t)$ is obtained applying the Bayes rule to the generative distributions:

$$p(\alpha_t \mid b_t) = \frac{p(b_t \mid \alpha_t)\, p(\alpha_t)}{p(b_t)}. \tag{5}$$

5 Application Example

Figure 2 some of the different views included in the Philips HomeLab dataset [23]. A living room is shown from different perspectives. The task for this dataset is to recognize when the user is doing different actions, such *reading*, *watching tv*, *walking*, *eating* or *drinking*.



(a) Camera 1 (b) Camera 2 (c) Camera 5

(d) Camera 7 (e) Camera 8

Fig. 2. Different views of the HomeLab dataset

The different actions are observed on most of the views displayed, so the view-invariant classifier can be trained with the action samples grabbed from all the views. Then, for each view, the context model would be learned. This way, for camera 1 it is learned that the action *walk* is mostly performed at the clear space and *sit* close to the *footstool* on the left. For camera 2, that it is very likely to observe the action *walk* on the clear space on the left and *eating*, *drinking* or such related actions close to the table on the right. For camera 5 and camera 7 that close to the sofas is very likely to observe the actions *reading* or *watching tv*. For camera 8 the result would be that *drinking* or *eating* are very likely to be observed in all the image, as the table takes up most of the space. Then, this acquired knowledge would be used to average the probabilities of the action labels most likely to be observed on new action samples.

6 Conclusions and Future Works

This paper has presented and human action recognition system decomposed in two different steps. First, a view-independent classifier, shared between the different views to analyze, provides an initial guess of the action being performed. Then, this estimation is averaged using view-dependent knowledge to make the

final decision on the performed action. This way, the view-independent module can be reused if a new view has to be analyzed, being only necessary to define the view-dependent knowledge.

The system has to be experimentally validated on future works, to quantify how much the use of the context information improves the accuracy of the action classification. We are currently working on obtaining a true view invariant classifier to add to the system. This is done exploiting restrictions derived from simultaneous views of the current action. Other issue to explore is the usage of action models learned from different scenarios, and quantify the plausibility of transferring them to scenarios previously unobserved.

Acknowledgement

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. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, New York (2006)
2. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 0, 1932–1939 (2009)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005, vol. 1, pp. 886–893. IEEE, Los Alamitos (2005)
4. Dotu, I., Van Hentenryck, P., Patricio, M., Berlanga, A., García, J., Molina, J.: Real-time tabu search for video tracking association. In: Principles and Practice of Constraint Programming-CP 2009, pp. 21–34 (2009)
5. Gomez-Romero, J., Patricio, M.A., Garcia, J., Molina, J.M.: Context-Based Reasoning Using Ontologies to Adapt Visual Tracking in Surveillance. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, pp. 226–231. IEEE, Los Alamitos (2009)
6. Gritai, A., Sheikh, Y., Shah, M.: On the use of anthropometry in the invariant analysis of human actions. In: Proceedings of the 17th International Conference on Pattern Recognition ICPR 2004, vol. 2, pp. 923–926 (2004)
7. Lewandowski, M., Makris, D., Nebel, J.C.: View and Style-Independent Action Manifolds for Human Activity Recognition. In: Computer Vision–ECCV, vol. 2010, pp. 547–560 (2010)
8. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2007, pp. 1–8. IEEE, Los Alamitos (2007)
9. Martínez-Contreras, F., Orrite-Uruñuela, C., Herrero-Jaraba, E., Ragheb, H., Velastin, S.A.: Recognizing Human Actions using Silhouette-based HMM. In: IEEE Conference on Advanced Video and Signal-based Surveillance, pp. 43–48 (2009)

10. Parameswaran, V., Chellappa, R.: View invariants for human action recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2 (2003)
11. 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)
12. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
13. Rao, C., Yilmaz, A., Shah, M.: View-invariant representation and recognition of actions. International Journal of Computer Vision 50(2), 203–226 (2002)
14. Richard, S., Kyle, P.: Viewpoint Manifolds for Action Recognition. EURASIP Journal on Image and Video Processing (2009)
15. Robertson, N., Reid, I.: A general method for human activity recognition in video. Computer Vision and Image Understanding 104(2-3), 232–248 (2006)
16. Sheikh, Y., Sheikh, M., Shah, M.: Exploring the space of a human action. In: Tenth IEEE International Conference on Computer Vision ICCV 2005, vol. 1 (2005)
17. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE, Los Alamitos (2002)
18. Tipping, M.E.: Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research 1, 211–244 (2001)
19. Tran, D., Sorokin, A., Forsyth, D.: Human activity recognition with metric learning. In: Proceedings of the 10th European Conference on Computer Vision: Part I, p. 561. Springer, Heidelberg (2008)
20. Turaga, P., Veeraraghavan, A., Chellappa, R.: Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008, pp. 1–8. IEEE, Los Alamitos (2008)
21. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3d exemplars. In: IEEE 11th International Conference on Computer Vision ICCV 2007, pp. 1–7. IEEE, Los Alamitos (2007)
22. 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)
23. Wu, C., Aghajan, H.: User-centric environment discovery with camera networks in smart homes. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 41(2), 375–383 (2011)