ROBUST HUMAN ACTION RECOGNITION USING HISTORY TRACE TEMPLATES

Georgios Goudelis, Konstantinos Karpouzis and Stefanos Kollias

Image, Video and Multimedia Systems Laboratory,
National Technical University of Athens.
9 Heroon Politechniou str., 15780, Athens, Greece

ABSTRACT

Due to the growing use of human action recognition in every day life applications, it has become one of the very hot topics in image analysis and pattern recognition. This paper presents a new feature extraction method for human action recognition. The method is based on the extraction of Trace transforms from binarized silhouettes, representing different stages of a single action period. A final history template composed from the above transforms, represents the whole sequence containing much of the valuable spatio-temporal information contained in a human action. The new method takes advantage of the natural specifications of the specific Trace transform, such as noise robustness, translation invariance and scalability easiness and produces effective, simple and fast created features. Classification experiments performed on KTH action database using Radial Basis Function (RBF) Kernel SVM, provided very competitive results indicating the potentials of the proposed technique.

1. INTRODUCTION

Observation and analysis of the human behavior is an open research topic for the last decade. Recognizing human actions in everyday life, is a very challenging task that finds application in a variety of fields such as automated crowd surveillance, shopping behavior analysis, automated sport analysis and others. One could state the problem as the ability of a system to automatically classify the action performed by a human person, given the action containing video.

Although the problem can be easily grasped, providing a solution to this problem is a daunting task that requires different apprach to several sub-problems. The challenge of the task rises from various factors that influence the recognition rate. Complicated backgrounds, illumination variations, camera stabilization and view angle are only a few of them.

If we had to classify human action recognition in different sub-categories, one could do it by taking into consideration the underlying features used to represent the various activities. As authors spot in [1], there are two main classes based on the underlying features representing activities. The most successful one is based on "dynamic features" and comprises the research object for the majority of current studies. The second one is based on "static pose based features" and provides the advantage of extracting features from still images.

A system inspired by Local Binary Patterns (LPB) is presented in [2] that is resilient to variations in textures by comparing nearby patches, while it is silent in the absence of motion. LPB based techniques have also been proposed in [3] where the space-time volume is sliced along the three axes (x, y, t) to construct LBP histograms of the xt and yt planes. Another approach in [4] in order to capture local characteristics in optical flow, computes a variant of LBP and represents actions as strings of local atoms. In [5] another approach, inspired by biology, uses hierarchically ordered spatio-temporal feature detectors. Space time interest points are used to represent and learn human action classes in [6]. Improvement of result has reported in [7], [8] where optical flow based information is combined with appearance information. In a newer study in [9], a spatiotemporal feature point detector is proposed, based on a computational model of saliency.

As mentioned earlier, the features used for human action recognition can be extracted either from video sequences or still images describing different static poses. The methods that use still images, are mostly silhouette based and although they do not present the accuracy of the sequence based techniques, they provide the main advantage of single frame decision extraction. Representative samples of this category are the methods presented in [10], [11]. More specifically, in [11] behavior classification is achieved extracting eigenshapes from single silhouettes using Principal Component Analysis. Modelling of human poses from individual frames in [10],uses a bag-of-rectangles method for action categorization.

Other technique in [12] involves infrared images to extract more clear poses. In following, classification is achieved using single poses based in Histogram of Oriented Gradients (HOG) descriptors. A type of HOG descriptors is also used in [13], on a set of predefined poses representing actions of hockey players. To better cope with articulated poses and cluttered background, authors in [14] extend HOG based descriptors and represent action classes by histograms of poses

This work was funded by ICT-Project Siren, under contract (FP7-ICT-258453).

primitives. Also in contrast to other techniques that use complex action representations, authors in [15] propose a method that relies on "key pose" extraction from action sequences. The method selects the most representative and discriminative poses from a set of candidates to effectively distinguish one pose from another.

The method we present in this paper, proposes the extraction of new simple features for human action recognition, based on the well known specifications of the Trace transform. In detail, in this study, we use the Radon form of Trace to create volume templates ($History\ Traces$) that each one represents a single period for every action. Radial Basis Function (RBF) kernel Support Vector Machine (SVM) used for the evaluation of the system, shows a very competitive performance of 87,7%.

The rest of the paper is organized as follows. An overview of the proposed system is given in 2. In 3 $History\ Trace$ templates creation is described. The experimental procedure is provided in 4 followed by conclusion and future work in section 5.

2. OVERVIEW OF THE PROPOSED SYSTEM

The most common way to capture a human action is by using a standard 2D camera. Thus, the action is contained in a video sequence comprised by a number of different frames. In our scheme, we work on KTH database which comprises a large number of action video sequences. Since background in all videos is uniform, we subtract it using a grassfire algorithm [16]. Silhouette extraction is a common technique in many different studies concerning observation of human dynamics [17], [18]. As it is used in most of the human action algorithm approaches, we constructed the testing and training examples manually, segmenting (both in space and in time). We have also aligned the provided sequences. This way, each action sample is represented by a time-scaled video that containes one period.

Although the background is uniform, extracted silhouettes appear to be noisy as there is still a number of external factors (such as illumination conditions etc.) that dramatically affect the result. However, due to trace transform specifications, the new features created, present to be robust to noise and do not require prior filtering. Thus, a trace transform is created for each silhouette. A final template that represents the entire movement, is created as the result of the integration of the binary transformations to it.

In following, the final templates comprise the vectors that will train equal to the number of classes, RBF kernel SVMs. Classification is achieved by measuring the distance of the test vector from the support vectors of each class. However, since the objective is to evaluate the overall performance of the new scheme, we measured the total number of correct classifications for every vector passing from each trained SVM respectively. For testing, we followed a leave-one-person-out protocol. Further details on the experimental procedure are provided in the corresponding section 4.

3. CONSTRUCTING HISTORY TRACE TEMPLATES

It has been shown [19] that the integrals along straight lines defined in the domain of a 2D function can fully reconstruct it. Trace transform is produced by tracing an image along with straight lines where certain functionals of the specific function are calculated. The result of Trace transform, is another 2D image which consists a new function that depends on the parameters (ϕ, p) that characterize each line. Definition of the above parameters for an image Tracing line is given in Figure 1. Different Trace transforms can be produced using different functionals. In this work, we choose the appropriate computation of the corresponding Trace functionals according to [19] so that we take advantage of noise robustness of Trace and invariability to translation, and scaling. The specific reconstruction is actually a subcase of the Trace, the so-called Radon transform which has found a variety of important applications, from computerized tomography to gait recognition [17].

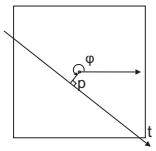

Fig. 1. Definition of the parameters of an image tracing line.

Let $f(x, y)$ be a 2D function in the Eucledean plane $\mathbf{R}^2$ taken from an action video sequence containing an extracted binary silhouette. The Trace Transform R_f, is a function defined on the space of straight lines L in $\mathbf{R}^2$ by the integral along each such line and is given by:

$$R_f(p,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(p - x\cos\theta - y\sin\theta)dxdy \quad (1)$$

where $R(p, \theta)$ is the line integral of the image along a line from $-\infty$ to ∞. p and θ are the parameters that define the position of the line. So, $R_f(p, \theta)$ is the result of the integration of f over the line $p = x\cos\theta + y\sin\theta$. The reference point is defined as the center of the silhouette.

As human actions are in fact spatio-temporal volumes, the aim is to represent as much of the dynamic and the structural information of the action as possible. At this point, Trace transform shows a greate suitability for this task. It transforms 2-dimensional images with lines into a domain of possible line parameters, where each line in the image will give a peak positioned at the corresponding line parameters. When Trace transform is calculated with respect to the center of the

silhouette, specific coefficients will have capture much of the energy of the silhouette. These coefficients will vary during time and will provide great differences from one action to another for the same time-frame.

Besides structural information, in order to also capture the temporal information included in a movement, we propose the construction of a $History\ Trace\ Template$. This template is actually a continuous transform in the temporal direction of a sequence. Let $f(p, \vartheta, t)$ be a human action sequence. If $\breve{g}_n(p, \theta)$ is the Trace transform of a silhouette $s_n(p, \theta)$, for the n frame where $n = 1...N$, then the history Trace Template for the action sequence will be given from:

$$T_N(p, \theta) = \sum_{n=1}^{N} \breve{g}_n(p, \theta). \quad (2)$$

This way the resulting features will be a function of multiple significant distinctions contained in multiple transforms produced for every action period respectively. As mentioned above, in our work all action periods have been timescaled to the same nuber of frames N. Figure 2 shows the transformations for each extracted silhouette received from one walking period. The final history Trace template is shown on the bottom side of the figure.

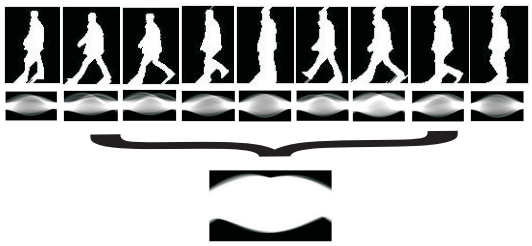

Fig. 2. Extracted silhouettes and Radon transforms for one walking period. History Trace template is shown on the bottom.

4. EXPERIMENTAL RESULTS

In this section, we will present the experimental results in order to demonstrate the efficiency of the proposed scheme for human action recognition. In our experiments, the leave-one-person-out cross-validation approach was used to evaluate the method performance. The experiments were performed on an Intel Core i5 (650@3,2 GHz) processor with 4GB RAM memory. For the experiments the KTH [20] action database was used. Samples from the dataset used for different type of action are illustrated in Figure 3.

The current video database contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios, under different illumination conditions:

Fig. 3. Action samples from KTH database for walking, jogging, running, boxing, hand waving and hand clapping respectively.

Table 1. Classification percentages (%) for the different action types of KTH database.

Action Type	Boxing	Handclapping	Handwaving	Jogging	Walking	Running	Overall
Classification	92.2	90.0	85.6	84.3	88.1	86.0	87.7

outdoors, outdoors with scale variation (camera zoom in and out), outdoors with different clothes and indoors. The database contains 600 sequences. All sequences were taken over homogeneous backgrounds with a static camera with 25fps frame rate.

In our experiments the sequences have been downsampled to the spatial resolution of 160*120 pixels and have a length of four seconds in average. The training examples were constructed by manually segmenting (both in space and in time) and aligning the available sequences. The background was removed using a grassfire algorithm [16]. The leave-one-person-out cross-validation approach was used to test the generalization performance of the classifiers for the action recognition problem.

At this point, we should note that human action recognition is a multiclass classification problem. We cope with this, by constructing the problem as a generalization of binary classification. More specifically we trained 6 different SVMs (one for each class) using an one-against-all protocol. The final decision was made by assigning each testing sample to a class $\mathcal{C}_a$, according to the distance d of the testing vector from the support vectors. Where $\mathcal{C}_a$ is the set of templates assigned to an action class (e.g boxing). However, since we wanted to evaluate the generalization of the algorithm in a more broad way, we measured the successful binary classifications of every sample, tested on each of the 6 different trained SVMs. This way we managed to produce 25*6*24=3600 classifications instead of 600 (persons*actions*samples per person). The results, indicated a very competitive classification rate of 87.7%. This becomes more interesting considering that the new method performs very fast. For 25 iterations (testing all samples), training included, it required 6 minutes while each sample was tested within 0.01 seconds. Classification rates for each class used in the experiments are given in Table1.

5. CONCLUSION

In this paper, a novel method for efficient feature extraction for human action recognition is presented. A subcase of Trace transform, the so-called Radon, is used to produce a number of templates from binary silhouettes. The extracted transforms are finally integrated in a $History\,Trace$ template that contains much of the spatio-temporal information of the action. Experiments performed in KTH database using RBF kernel SVMs show a very competitive 87.7% classification rate and proved that the method is able to efficiently distinguish very similar classes (e.g. jogging vs running). It is worth noting that the new method proved to be very robust in illumination variations, noise and scaling (zoom-in zoom-out) conditions.

At this point we should mention that work is under progress. Future work will examine all specifications for different Trace functionals and will show how they can improve classification rates. Different classifiers will be compared, while there is also an effort to integrate the concatenation of each action to the final template.

6. REFERENCES

[1] Christian Thurau and Va'clav Hlava'c, "Pose primitive based human action recognition in videos or still images.," in *CVPR*. 2008, p. 8, IEEE Computer Society, Madison, USA, Omnipress.

[2] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *(ICCV)*. IEEE, 2009.

[3] V. Kellokumpu, G.Y. Zhao, and M. Pietikainen, "Human activity recognition using a dynamic texture based method," in *BMVC08*, 2008.

[4] C.J. Yang, Y.L. Guo, H.S. Sawhney, and R.T. Kumar, "Learning actions using robust string kernels," in *HUMO07*. Springer, 2007, pp. 313–327.

[5] Hueihan Jhuang, Thomas Serre, Lior Wolf, and Tomaso Poggio, "A biologically inspired system for action recognition," in *ICCV*. 2007, pp. 1–8, IEEE.

[6] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," in *(IJCV)*, 2008.

[7] K. Schindler and L.J. Van Gool, "Action snippets: How many frames does human action recognition require?," in *CVPR08*, 2008, pp. 1–8.

[8] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Conference on Computer Vision & Pattern Recognition (CVPR)*. 2008, IEEE.

[9] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *(CVPR)*, 2009.

[10] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, "Behavior classification by eigendecomposition of periodic motions," in *Pattern Recognition*, 2005, pp. 38:1033–1043.

[11] N. Ikizler and P. Duygulu, "Human action recognition using distribution of oriented rectangular patches," in *(ICCV)*. 2007, pp. 271–284, Human Motion.

[12] L. Zhang, B. Wu, and R. Nevatia, "Detection and tracking of multiple humans with extensive pose articulation," in *(ICCV)*, 2007, pp. 1–8.

[13] Wei-Lwun Lu and James J. Little, "Simultaneous tracking and action recognition using the pca-hog descriptor," in *(CRV)*. 2006, p. 6, IEEE Computer Society.

[14] Thurau C. and Hlavac V., "Pose primitive based human action recognition in videos or still images," in *CVPR*. 2008, pp. 1–8, IEEE Computer Society.

[15] Sermetcan Baysal, Mehmet Can Kurt, and Pinar Duygulu, "Recognizing human actions using key poses," *Pattern Recognition, International Conference on*, pp. 1727–1730, 2010.

[16] Georgios Goudelis, Anastasios Tefas, and Ioannis Pitas, "Automated facial pose extraction from video sequences based on mutual information," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 18, no. 3, pp. 418–424, 2008.

[17] Nikolaos V. Boulgouris and Zhiwei X. Chi, "Gait recognition using radon transform and linear discriminant analysis," *IEEE Transactions on Image Processing*, vol. 16, no. 3, pp. 731–740, 2007.

[18] Irene Kotsia and Ioannis Patras, "Relative margin support tensor machines for gait and action recognition," in *CIVR*, 2010, pp. 446–453.

[19] Alexander Kadyrov and Maria Petrou, "The trace transform and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, pp. 811–828, August 2001.

[20] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: A local svm approach," in *In Proc. ICPR*, 2004, pp. 32–36.