Human Activity Recognition using Local Shape Descriptors

Sharath Venkatesha, Matthew Turk
Department of Computer Science, University of California, Santa Barbara
{sharath, mturk} @cs.ucsb.edu

Abstract

We propose a method for human activity recognition in videos, based on shape analysis. We define local shape descriptors for interest points on the detected contour of the human action and build an action descriptor using a Bag of Features method. We also use the temporal relation among matching interest points across successive video frames. Further, an SVM is trained on these action descriptors to classify the activity in the scene. The method is invariant to the length of the video sequence, and hence it is suitable in online activity recognition. We have demonstrated the results on an action database consisting of nine actions like walk, jump, bend, etc., by twenty people, in indoor and outdoor scenarios. The proposed method achieves an accuracy of 87%, and is comparable to other state-of-the-art methods.

1. Introduction

The aim of activity recognition is to recognize the actions and goals of one or more agents, based on observations of the agents' actions and the environmental condition. Activity recognition has promising applications in areas of surveillance, human performance analysis, computer-human interfaces, content-based image retrieval/storage and virtual reality. For example, there is a need for intelligent security systems as large number of surveillance cameras are being deployed in public spaces. As the number of cameras grows exponentially, it is difficult for an operator to monitor the video streams for extended periods of time. This also requires tremendous amounts of data to be stored for analysis. Instead we need systems which are able to automatically detect, categorize, and recognize human activity, calling for human attention only when necessary. An online activity recognition system is the required solution.

We propose a method for online activity recognition that is invariant to the length of the video sequence. Instead of storing the video data and processing the frames offline, the method directly learns the activity in the scene and is able to categorize it. Only a few seconds of the input video which contains the activity is required for analysis. We claim that the shape involved in the person's activity can be used to build a signature which is representative of the action. Using local shape descriptors defined for interest points on the shape boundary we build an *Action Descriptor*, a histogram with bins corresponding to the representative descriptor set. An SVM then classifies the activity to one of the several known common actions used in the training phase.

Related Work: A variety of descriptors has been defined on the 2D shape and the 3D volume of an object to analyze its activity. Spatio-Temporal Volume (STV) is a widely used methodology for activity recognition. Gorelick et al. [1] define Poisson based shape descriptors on the STV of an action and use nearest neighbor classification to recognize the activity. Similarly, Yanke et al. [3] match volumetric representation of an activity to an over-segmented STV of the scene. Traditional methods also include using a spatio-temporal pattern like Motion History Image (MHI) as defined by Bobick et al. [5].

On the other hand, studies in the field of object recognition in 2D images [5] have demonstrated that silhouettes contain detailed information about the shape of objects. Our work has similarity to work by Schuldt et al. [4] who define Gaussian jet based flow descriptors for the motion in the sequence and use an SVM for classification.

Problem Definition: We address the problem of identifying the actions/activity of a person in the video. A set of nine actions like walk, bend, jump, run, etc., is analyzed. These form a subset of basic human actions performed as part of daily activity. Many complex actions are composed of actions from this basic set in sequence or in parallel. For example, a suspicious activity like leaving a bag in a public place may have the suspect walk, bend and run in a sequence. We propose a method which classifies these actions independently with reasonable accuracy; the method can be extended to recognize complex activities as a sequence of basic activities.

2. Silhouettes and Interest Point Selection

Different techniques for background subtraction have to be adopted depending on the complexity of the scene, movement of the camera and illumination changes. The clothes worn by a subject and the degree of movement involved can result in silhouette images being extremely noisy. We analyze the shape of the moving foreground object and hence need to extract a reasonable silhouette by this process.

We acknowledge the fact that background subtraction is difficult and have considered a simple scenario where the camera is fixed. This is valid as surveillance applications generally use fixed position cameras. For an outdoor scenario, we model the background as an *adaptive Mixture of Gaussians*. For indoor scenarios we use a simpler *median filter* method. Further, we smooth the contour boundary to remove noisy edges.

Figure 1. A sample frame in a run sequence, its extracted silhouette, and interest points on contour

The boundary of the silhouette is traced to extract the contour. Interest points are detected on the contour considering two aspects – points of high curvature are preferred, and points are evenly spread along the contour. We consider a maximum of 30 points on the contour but this number varies depending on its shape. Fig. 1 shows a sample frame, the silhouette and its contour with the interest points marked.

3. Local Shape Descriptor

A *local shape descriptor* of a feature point depends only on its neighboring points and captures local characteristics. Local shape descriptors are robust to partial occlusions and usually insensitive to global shape transformations such as rotation, articulation, or view changes. Such features can be adapted to the size and frequency of moving patterns; this results in a video representation that is stable with respect to corresponding transformations.

Each of the interest points selected on the contour is described by two measurement based shape descriptors: *Turning angle* and *Distance across the shape*. As shown in Fig. 2, for any feature point *p*, its immediate predecessor *q*, and its successor *r*, *qpr* forms a turning angle (TA). If the interior bisector of *qpr*

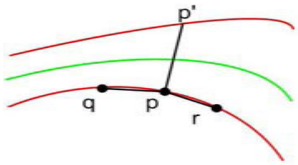


Figure 2. Shape descriptors: TA (∠ qpr) and DAS (pp′)

intersects the contour at point *p'*, the length of *pp'* is the distance across the shape (DAS) at point *p*. If the bisector intersects the shape multiple times, the distance to the closest intersection is used. The descriptors are translation and rotation invariant and are made scale invariant by normalizing with respect to the size of the shape. These descriptors lead to a 16 length feature vector for each point.

4. Temporal Information

To capture the change in shape over time, we track the object contours across successive frames and in particular the interest points detected on the contour. Interest points may be lost due to noise or occlusions. Hence, a one-to-one matching between interest points does not exist. A partial shape matching method can be used to match interest points which are found in common across the frames. We use the method by Chen et al. [2].

For any pair of matched points, the difference in *[x, y]* location of the points, normalized with respect to the size of the image forms the temporal component of the descriptor. The difference in centroid location of the shape in the two frames is also noted and it forms the temporal component of an interest point, if its match is not found in the next frame. The temporal information is added to the feature vector for each point.

5. Using the Bag of Features

Our objective is to obtain a set of representative shape descriptors which can be used to build the action descriptor. *Bag of Features* is a well established technique for the visual classification of objects, categories of objects and textures. The images are represented by a collection of regions ignoring their spatial relationships. This concept is used for our purpose – we consider the descriptor vectors for all the interest points ignoring their spatial positions. In general, let us consider that we have *m* different actions and *k* sequences of each action in our training set. Each sequence may have an average of *p* frames and *q* feature points selected per frame for which the descriptors have been calculated. Hence the bag of features will have $N = m * k * p * q$ descriptors.

The resulting distribution of descriptors in descriptor space has to be quantized. We use *k-means* clustering with *cosine distance [u, v]* as a metric, which is the angular cosine distance between vectors $\underline{u}$ and $\underline{v}$. This metric is more suitable than the common Euclidean measure, as our descriptors are based on shape measurements. We chose to keep the complexity low by selecting the number of clusters N to be 100. Experiments also showed that increasing the number of clusters does not improve the efficiency of the method.

6. The Action Descriptor

For a given action sequence, we evaluate descriptor vectors for every frame in the input video and consider a bin of size N. The action descriptor is represented by a histogram, where a vote for the closest matching descriptor (measured by a distance function) in N is considered for all the descriptors. The histogram is then normalized to provide invariance to the number of frames *(p)*, and the number of descriptors calculated per frame *(q)* of the input video.

We define a distance function which considers the matching measures *TA* and *DAS* independently. For two feature points *a* and *b* it is defined as

$$C_{DAS}(a, b) = \frac{|DAS(a) - DAS(b)|}{DAS(a) + DAS(b) + \varepsilon} \quad (1)$$

$$C_{TA}(a, b) = \frac{|TA(a) - TA(b)|}{TA(a) + TA(b) + \varepsilon} \quad (2)$$

where DAS (*) and TA (*) are the DAS and TA feature vectors of the feature point respectively, and ε is a small constant to make the denominators nonzero. The similarity function Φ is defined as

$$\Phi(a, b) = W(K_{DAS} - C_{DAS}(a, b)) + (1 - W)(K_{TA} - C_{TA}(a, b))$$

where *W* and *(1 - W)* are weights given to the TA and DAS descriptors, and K_{DAS}, K_{TA} are positive constants. In our experiments, W= 0.7 is used. K_{DAS} and K_{TA} are constants used to change the sign of the value obtained as similarity measure.

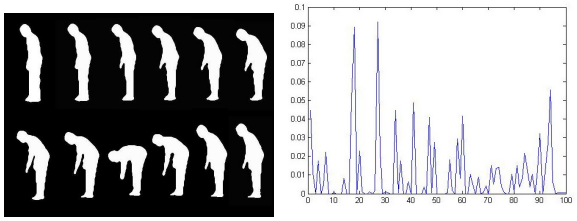
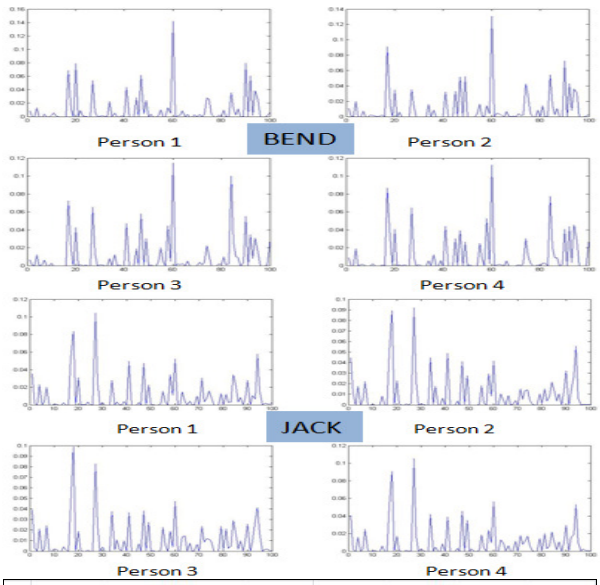


Figure 3. Sample silhouettes from a sequence of *bend* action and the corresponding *Action Descriptor* with a histogram bin size of 100.

For the *Action Descriptor* to be discernable, the intra class distance (i.e., among the histograms belonging to the same action of different people) should be less, compared to the inter class distance (i.e., among the histograms of different actions of the same person). Figure 4 shows histograms of two actions *bend* and *jack* for four people. We see that the histograms of an action are almost similar for all the people. This was verified by the *Chi-Square* distance between the histograms. As shown in the table above, the closest match belongs to the same action class.



	BEND				JACK			
B	1.000	0.081	0.083	0.084	0.113	0.108	0.119	0.135
E	0.081	1.000	0.087	0.073	0.168	0.168	0.155	0.169
N	0.083	0.087	1.000	0.055	0.124	0.118	0.136	0.134
D	0.084	0.073	0.055	1.000	0.152	0.147	0.155	0.158
J	0.113	0.168	0.124	0.152	1.000	0.052	0.090	0.098
A	0.108	0.168	0.118	0.147	0.052	1.000	0.074	0.086
C	0.119	0.155	0.136	0.155	0.090	0.074	1.000	0.037
K	0.135	0.169	0.134	0.158	0.098	0.086	0.037	1.000

Figure 4. Sample *Action Descriptors* for actions *bend* and *jack* for four subjects. The table shows chi-square distances between the histograms with the least distance highlighted in red, which belongs to the same action class.

7. Experiments and Results

The action descriptors were trained using a multi-class SVM. For the experiments the kernel type used was a radial basis function (RBF) with degree 3.

For the evaluation, we recorded a video database containing nine types of human actions namely

actions = [bend, wave-both-hands, jump, jumping-jack, wave-one-hand, hop, run, side-walk, walk]

of 12 subjects in an indoor scenario. We have also used dataset made available from Gorelick et al. [1] which consists of sequences of the same set of actions performed by 9 subjects in an outdoor scenario.

The sequences consider homogeneous backgrounds with a static camera at 25 fps. The sequences are of varied lengths (from 2-10 seconds) and are of standard and low resolutions (640x480 and 180x144 pixels). We scale the frames from [1] by a factor of 3, so that the number of feature points extracted on the contour remains approximately the same across the full dataset. In total the dataset contains in excess of 250 sequences. Fig. 5 shows samples from the dataset.

Figure 5. Sample frames from the indoor video database

The sequences were divided with respect to the subjects into a training set (8-12 persons) and a test set (8-12 persons) for different experiments.

		PREDICTED CLASSES								
		BEND	WAVE -2	HOP	JACK	JUMP	WAVE -1	RUN	SIDE-WALK	WALK
ACTUAL CLASSES	BEND	10								2 [E]
	WAVE-2		9		3 *					
	HOP			12						
	JACK				12					
	JUMP					11			1 [E]	
	WAVE-1				4 * [E]		8			
	RUN							12		
	SIDE-WALK			2 [E]					10	
	WALK							2 *		10

Figure 6. Confusion matrix of classification results

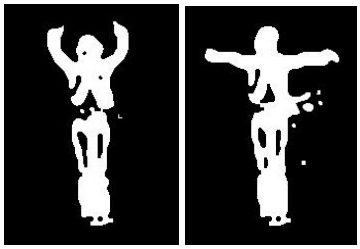


Figure 7. Erroneous silhouettes for Wave-2 and Jack

The results of the classification are shown as a confusion matrix in Fig. 6. The accuracy achieved is around 87%. The table displays the number of sequences classified correctly in the diagonal, and [E] or [*] for the sequences which are classified wrongly. [*] is marked for the misclassified actions which have similar shaped silhouettes. Actions like walk and run, wave-both-hands/wave-one-hand and jumping-jack have similar sub-actions. The method also assumes that the input silhouettes are reasonable and they depict the shape of the action in the particular frame. The analysis of erroneous classifications shows that the background subtraction has failed (marked by [E]). Fig. 7 shows two examples for which background subtraction failed.

8. Discussion

We provide a qualitative comparison with two similar methods. Gorelick et al. [1] achieve an accuracy of 97.8% on a set of around 80 video sequences consisting of 9 actions. Schuldt et al. [4] consider a dataset of more than 2000 video sequences and report an accuracy of around 70%. Our dataset is more generic than that of [1], though we do not consider videos with different orientations as in [4]. Further, our method builds a descriptor from an online video stream to provide the results of activity recognition. This is a clear advantage compared to existing methods.

9. Robustness to changes in Pose

We experimented with data where the subject was captured at angles 30^0 and 45^0 from the frontal direction. The results show that the proposed method is consistent to changes in pose up to 45^0; Above this angle, the silhouettes do not capture the shape of an action resulting in failure of the method. We plan to investigate this in our future work.

10. Conclusion

The proposed method requires no storage of frames and is able to detect the actions from a continuous video stream. The performance of the method depends on a reasonable background subtraction technique and is shown to be comparable to other state of art methods. The use of local shape descriptors also provides robustness to occlusions and view changes. The method has ample scope to be integrated into surveillance suites for activity recognition. The method can also be extended to detect complex human actions.

References

1. L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, “Actions as space-time shapes”, PAMI 2007.
2. L. Chen, R. Feris and M. Turk, “Efficient partial shape matching using Smith-Waterman algorithm”, CVPR Workshops, 2008.
3. Y. Ke, R. Sukthankar and M. Hebert, “Event detection in crowded videos”, ICCV, 2007.
4. C. Schuldt, I. Laptev and B. Caputo, “Recognizing human actions - a local SVM approach”, ICPR 2004.
5. A. Bobick and J. Davis “The recognition of human movement using temporal templates”, PAMI 2001.