Action Energy Images for Reliable Human Action Recognition

Varsha H Chandrashekhar, K S Venkatesh
Department of Electrical Engineering,
Indian Institute of Technology Kanpur
Kanpur – 208016 (INDIA)
venkats@iitk.ac.in

Abstract: *We present an approach for Human Activity Recognition using a compact 2D spatio-temporal action representation called Action Energy Image (AEI). Our hypothesis is that the AEI carries useful structure and gross motion information which is sufficient for activity classification. We construct the Eigen Activity Space by performing PCA on AEIs of various activities and use it for recognition of a test activity sample. The promising results obtained by our method demonstrates the capacity of AEI to discriminate human actions. Our method is robust to anthropometric changes of actors and changes in action speed. It is also invariant to view point changes to a small extent.est.*

Keywords: human action recognition. silhouette extraction, view invariance, eigen action space.

1 Introduction

Human action recognition is a very important component of surveillance systems for event based analysis of surveillance videos. Human activity analysis in surveillance scenarios involves detection of abnormal human actions which are deviant from normal human activities. For example in a shopping mall where people normally walk from one counter to other running activity may be defined as an abnormal action and could be an event of interest for surveillance purpose. Analysing human action is particularly challenging problem due to complex non rigid and self occluding motion of the articulated human body with its large degrees of freedom and many sources of variability in actions owing to changes in viewpoint, anthropometry, dress, execution rate, individual styles etc. It has many other interesting applications like video indexing and browsing, motion analysis for medical purposes, sports video analysis, choreography, HCI etc.

Most of the recent action recognition approaches can be divided into the following categories: 2D methods, 3D spatio-temporal methods, and 2D methods using 3D constraints. 2D methods describe an action as a sequence of poses. 2D Appearance based methods describe each pose by 2D image of raw gray scale, body contours or edges [11], color distribution, wavelet responses, background subtracted silhouettes [8,20] etc. These methods are sensitive to the changes in camera view angle and the changes in appearance, which is not preserved across different clothing. Amongst the 2D approaches using motion based gradients, optical flow is most popularly used [5, 8, 10, 19]. Optical flow images are least affected by appearance but are too noisy due to the inaccuracies involved in flow computation. A few approaches like [8] use both appearance and motion based features. Both structural and dynamic characteristics of human activities are important for activity recognition, [20].

[1, 15, 17, 21] analyse 3D S-T volume(STV) for activity recognition. In [1] Blank et al. use solution to Poison equation to determine space time saliency features of the action volume. In [21] Yilmaz and Shah use differential geometric properties of action STV.By looking at the space time shape only they ignore the intensity information inside the shape. In [15] Schuldt et al. use S-T gradients to _nd S-T interest points for activity recognition using SVM and in [17]Shechtman and Irani use S-T gradients to determine correlation between ST patches of two different videos. However, gradient based methods are unreliable for low quality videos, motion discontinuities and motion aliasing.

Seitz and Dyer [16] developed af_ne invariant matching for cyclic activities using 3D constraints on 2D image measurement data. Later it was extended by many authors for view invariant matching of non periodic actions. They use epipolar geometry rules to _nd bounds on image observation matrix to perform recognition. Most of these approaches model action using motion trajectory [12, 13] or using a set of body joints [6, 9, 18]. However, representation of complex human action by the trajectory of a single point or a few points is limited and ambiguous while joint based models require extremely accurate segmentation that is difficult to achieve in practice.

The rest of the paper is organized as follows. We explain our motivation and situate our work in the context of previous work in section 2. In section 3 we explain the steps followed in the construction of the Action Energy Image and its properties, We discuss construction of activity space in section 4. Experiments and results are presented in section 5, and conclude the discussion in section 6.

2 Motivation and Related Work

2.1 Motivation We ask the question whether from a single photograph of a scene, it is possible to derive information about who is doing what. For example, consider Figure 1, at _rst sight one can interpret that the man in the scene is running. The capacity of human vision system to recognize the nature of the movement from a single snap shot taken at suitable time inspires us to study average gross motion templates of the performer unlike the complex 3D approaches [1,15,17,21] which deal with the XYT shape volume, created by the actor as he performs the action or the sequence matching approaches [11,17,20] where features from individual(key) poses of the activity sequences are matched. We therefore evaluate the possibility of using a DC stance image of an activity as an

activity descriptor and we demonstrate that this image is sufficiently informative for activity recognition purposes.

2.2 Related Work Bobick and Davis, Rosales and Scarloff [2,14] used 2D temporal templates MEI and MHI for action recognition: however, the coarse representation of templates using Hu moments is insufficient to discriminate similar actions. We use eigen decomposition of an AEI in eigen activity space obtained by PCA, which best represents the AEI data in least-square sense. Most human actions are repetitive over time, so modeling the recency of motion using the MHI is not very meaningful for periodic actions like walking, running etc. MHI also depends on length and position of the time window over which MHI is computed. This is not an issue with AEIs as they are computed by averaging silhouettes. Unlike the MEI which captures only 'where the motion occurred' the AEI captures 'where and how much motion occurred'. MEIs carry less structural information since they are computed by accumulating motion images obtained by image differencing while the AEI carries information about both structure and motion. Our work is closely related to [7] in which the authors represent human gait a by gait energy image (GEI) to solve the tougher problem of gait identification. The authors show the inherent representational power of GEI and demonstrate that matching features from real gait templates achieves better performance than direct matching between individual silhouette frames. In [11], action is represented by a stream of successive poses forming action trajectories in eigen space. It employs blurred images to get rid of the dress problem and constructs a tuned eigen space by taking mean of similar postures of different persons to achieve body shape invariance. However they don't address the problem of speed variation. Our approach represents an action by a single point in Eigen Activity Space. And since it is template based it is robust to changes in speed of the activity. There is no need of time alignment of activity frames as in case of approaches based on matching features from key poses of an activity. We also compute the mean of the points corresponding to different subjects in activity space to represent average action point for every action. This makes our approach body shape invariant. On a conjecture similar to the one proven in [18], that the variability in AEIs of the same action resulting from various factors can be captured by the eigen action bases provided that the training data has exemplars spanning the complete action space, we decompose the AEI of a test action sample along action bases obtained by eigen analysis. The comparable results obtained by our method using simple minimum Euclidean distance classifier demonstrates our claim that the AEI of an action contains adequate information of structure and motion for recognition purposes.

3 Feature Extraction

3.1 Background subtraction/Silhouette extraction: We use a GMM background model similar to the one described in [3]. The recent history of each pixel $X_1, \ldots, X_t$ is modeled by a mixture of K Gaussian distributions. The probability of observing the current pixel value is

$$P(X_t) = \sum_{i=1}^{K} w_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}),$$

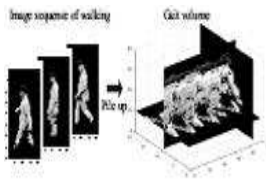
where K is the number of the distributions, $w_{i,t}$ is an estimate of the weight of the i th Gaussian in the mixture at time t, $\mu_{i,t}$ the mean value, and $\Sigma_{i,t}$ the covariance matrix of the i th Gaussian of the mixture at time t, where η is a Gaussian probability density function

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma^{1/2}|} \exp^{-\frac{1}{2}(X_t-\mu_t)^T \Sigma^{-1}(X_t-\mu_t)}$$

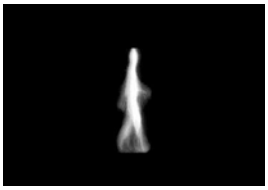
We also perform shadow detection using the shadow mask given in [4]. Figure 2(b) shows our background subtraction results after shadow removal. We perform silhouette normalization and nullify the translational component of motion by piling the silhouettes in 3D ST volume as shown in Figure 2(c). However, we compute and save the translational components of velocity V_x, V_y of the centroid of silhouette bounding box, which we use later on for improving our recognition performance.

(a) Original video

(b) after background subtraction



(c) 3D spatio temporal volume activity



(d) Activity Energy Image

Figure 2: Background Subtraction and Feature Extraction

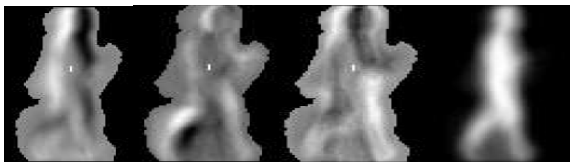
3.2 AEI representation: We consider pixel intensity at each pixel in XY direction of the ST volume as a 1D function of time $f_{xy}(n)$ and perform 1D Fourier analysis of this signal at each xy.

$$F_{xy}(k) = \sum_{n=1}^{N} f_{xy}(n) * \exp^{-\frac{j*2*\pi*k*n}{N}}$$

where N is number of frames over which activity analysis is carried out in order to determine what activity was performed by the agent in the last N frames. We construct a 2D AEI by taking $F_{xy}(0)$ for all xy as shown in Figure 2(d). AEI is a compact 2D representation of average 3D spatio-temporal information. AEI is also computationally efficient. Consider Figure 2(d), gray level values of the pixels in the regions of legs and hands swing are the measures of frequency of motion of limbs occurring at those points. While the white pixels in the torso, head regions indicate the overall structure of the actor and average pose during the action performance. Thus, AEI captures both structural

and motion characteristics of an action. AEI is obtained by averaging operation so it reduces the noise effects of background subtracted noisy silhouettes.

4 Training and Classification



(a) (b) (c) (d)

Figure 3: PCA: (a), (b), (c) eigen activities e_0, e_1, e_2 respectively; (d) average image

4.1 Construction of the Activity Space: We perform PCA on the AEIs obtained from the training dataset (from [1]) consisting of 9 activities performed by 9 subjects. Each $K = M*N$ AEI is vectorized into a 1D column vector by stacking together pixel rows of the image one after the other. AEI of the m th action of n th subject is given by $\text{AEI}^{m,n} = [p_1, p_2, \dots, p_k]^T$ where $m,n = 1,2, \dots, 9$. Next we form a data matrix with these vectorized AEIs as its columns.

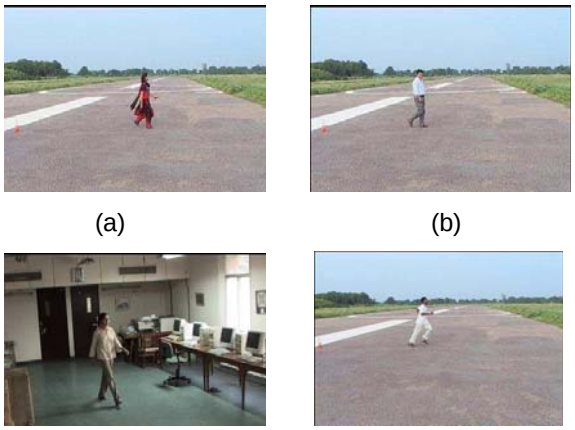
$$X = [\text{AEI}^{1,1}, \dots, \text{AEI}^{1,9}, \text{AEI}^{2,1}, \dots, \text{AEI}^{2,9}, \dots, \text{AEI}^{9,1}, \dots, \text{AEI}^{9,9}]$$

Covariance matrix of datamatrix X is given by $C = E[XX^T]$. Let the eigen vectors corresponding to the L highest eigen values $\lambda_0, \lambda_1, \dots, \lambda_{L-1}$ of C be given by $e_0, e_1, \dots, e_{L-1}$ where $1 << L << K$. The L dimensional space defined by e_l, $l = 0, 1, \dots, L$ is called the Activity Space. Figure 5 shows the Activity Space of 2 activities Walk and Run. For the sake of simplicity only 3 eigen vectors corresponding to the three greatest eigenvalues are shown. The corresponding eigen activity images are shown in Figures 3(a), 3(b) and 3(c). Figure 3(d) shows average image computed during PCA. We project the AEI of each training sample in this space. Figure 5 shows two clearly separable clusters of Walk and Run samples. Decomposition coefficients of $\text{AEI}^{m,n}$ are given by $d^{m,n} = [e_0, e_1, \dots, e_{L-1}]^T \text{AEI}^{m,n}$.

Each activity is represented by a single point in the Activity Space. Coordinates of an activity performed by different subjects are averaged to give a mean decomposition coefficient of the activity in the Activity Space. This reduces the effect of variability introduced by anthropometry and styles of actors on our approach: $d^m = (1/H)\Sigma^H_{n=1} d^{m,n}$. where H is the number of subjects. Given a test sample, its AEI is computed and projected in Activity Space. Recognition is carried out by Minimum (Euclidean) Distance Rule. Distance of the AEI of the test sample is calculated from the average points of all the action clusters. The test sample is recognized as activity n if $n = argmin_m\ dist(d^{test}, d^m)$.

5 Experiments and Results

We have used the database used by Blank et al. in [1]. The database consists of 9 actions by 9 subjects. These 9 actions are walking, running, jumping jack, jumping forward on two legs, jumping in place on two legs, galloping sideways, waving two hands, waving one hand and bending. We have performed leave one out experiment on this data set, as described in [1], i.e at a time we consider a video sequence of a subject performing certain action as a test sample and use rest of the 80 sequences for training.



(a) (b) (c) (d)

Figure 4: Our Data Set

We repeat this to recognize all the 81 video sequences ensuring every time that the sequence being tested is not a part of the training set. 5 out of 81 action videos were misclassified. 3 out of these 5 are the cases of extremely poor silhouettes extraction in which limbs, head etc. are missing. We have obtained error rate of 6.172%. Figure 6 shows the confusion matrix for this experiment. The diagonal numbers are the fractions of correct classifications and the off diagonal numbers are the fractions of misclassifications. We then used translational information V_x, V_y (Section 3.1) to improve our recognition. As seen in Figure 6 the confusion created in the recognition of A2, A7 and A8 can be completely resolved using the above information. This has reduced our error rate to 1.23% which, though higher than the error rate 0.38% achieved by [1] using space time shape information, is still much lower than 6.38% that is achieved by spatial-per-frame approach as cited in [1] in light of the simplicity of our approach.

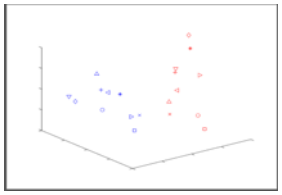


Figure 5

	A0	A1	A2	A3	A4	A5	A6	A7	A8
A0	1								
A1		1							
A2			0.89			0.11			
A3				1					
A4					1				
A5						0.89	0.11		
A6							1		
A7				0.22				0.78	
A8				0.11					0.89

Figure 6

Figure 5: 3D Activity Space for Walk-Run Classification

Figure 6: Confusion Matrix for 9 actions: A0-Walk, A1-Run, A2-Forward Jump, A3-Side Gallop, A4-Bend, A5-1 Hand Wave, A6-2 Hands Wave, A7-Jump in Place, A8-Jumping Jack

5.1 Analysis of Robustness: In this experiment we have used walking-running data from database in [1] for training and used our own database (Figure 4) for testing. Our dataset has samples of activity performed in inclined direction Figure 4(c), challenging dress styles like fluttering cloth Figure 4(a), and varied body postures shot on different backgrounds. We included a few inclined direction samples and a few samples with different dress styles with our earlier training database, [1] and tested the performance of our system on the rest of the samples of our database. We achieved 100% recognition rate in this experiment. This shows that our method is robust to some extent to body posture, dress and view point changes.

6 Conclusion and Future Work

Here, we have proposed a very simple and computationally efficient approach for action recognition using AEI. Our hypothesis is that the AEI captures structural and average motion information of an activity. We demonstrate that by using the AEI as a feature of action, separable action clusters in eigen activity space can be obtained. 100% recognition obtained for very similar actions like walk and run where most other approaches produce confusion clearly supports our hypothesis. Being template based, our approach is invariant to speed changes. We achieve anthropometric invariance by computing average action point for every action cluster in the activity space. Our approach could be made more robust to view angle changes by incorporating more samples of activities shot from different camera view points in our training database. Also, since this is an appearance based approach AEI representation is affected by different dress styles. Future work would involve making our approach robust to view angle changes and handling more challenging dress problems like occluded legs by obtaining internal contours from the silhouettes.

References

1 Blank, M., et al., "Actions as space time-shapes." *IEEE International Conference on Computer Vision, ICCV*, volume 2, pp. 1395-1402, Oct 2005.

2 Bobick, A., F., Davis, J., W., "The recognition of human movement using temporal templates." *IEEE Trans. Pattern Analysis Machine Intelligence, PAMI*, 23(3):257-267, Mar 2001.

3 Stauffer, C., Grimson W., E., L., "Adaptive background mixture models for real-time tracking." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, volume 2, pp. 252, June 1999.

4 Cucchiara, R., et al., "Improving shadow suppression in moving object detection with hsv color information." *IEEE Conf. Intelligent Transportation Systems*, pp. 334-339, Aug 2001.

5 Efros, A., A., et al, "Recognizing action at a distance." *IEEE International Conference on Computer Vision, ICCV*, vol. 2, pp. 726-733, Nice, France, Oct 2003.

6 Gritai, A., Sheikh, Y., Shah, M., "On the use of anthropometry in the invariant analysis of human actions." I*nt'l. Conf. on Pattern Recognition,ICPR*, vol 2, pp. 923- 926, Aug 2004.

7 Ju Han and Bin Bhanu. "Individual recognition using gait energy image." *IEEE Trans. Pattern Analysis Machine Intelligence, PAMI*, 28(2): pp.316-322, Feb 2006.

8 Niu, F., Mottaleb, M., "View invariant human activity recognition based on shape and motion features." *Int'l. Symp. on Multimedia Software Engineering, ISMSE*, pp. 546- 556, Dec 2004.

9 Parameswaran, V., Chellappa, R., "View invariants for human action recognition." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 2, pp. II 613-619, June 2003.

10 Polana, R., Nelson, R., "Low level recognition of human motion (or how to get your man without finding his body parts)." *IEEE Computer Society Workshop on Motion of Nonrigid and Articulate Objects*, pp. 77-82, Oct 1994.

11 Rahman, M., M., Ishikawa, S., "Robust appearance-based human action recognition." *Int'l Conf. on Pattern Recognition, ICPR*, vol. 3, pp. 165-168, Aug 2004.

12 Rao, C., et al., "View-invariant alignment and matching of video sequences." *IEEE International Conference on Computer Vision, ICCV*, vol. 2, pp. 939-945, Oct 2003.

13 Rao, C., Shah, M., "View invariance in action recognition." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 2, pp. 316-322, 2001.

14 Rosales, R., Sclaroff. S., "3D trajectory recovery for tracking multiple objects and trajectory-guided recognition of actions." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 2, pp. 123, 1999.

15 Schuldt, C., Laptev, I., Caputo, B., "Recognizing human actions: A local svm approach." *Int'l Conf. on Pattern Recognition, ICPR*, vol. 3, pp. 32-36, Aug 2004.

16 Seitz, S., Dyer, C., "View invariant analysis of cyclic motion." *Int'l Journal of Computer Vision, IJCV*, vol. 25, pp. 231-251, 1997.

17 Shechtman, E., Irani, M., "Space-time behavior based correlation." *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, CVPR*, vol. 1, pp. 405-412, June 2005.

18 Sheikh, Y., et al., "Exploring the space of a human action." *IEEE Int'l. Conf. on Computer Vision, ICCV*, vol. 1, pp. 144-149, Oct 2005.

19 Sidenbladh, H., "Detecting human motion with support vector machines." *Int'l. Conf. on Pattern Recognition, ICPR*, vol. 2, pp. 188 . 191, Aug. 2004.

20 Veeraraghavan, V., et al., "Role of shape and kinematics in human movement analysis." *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, CVPR*, vol. 1, pp. 730-737, July 2004.

21 Yilmaz, A., Shah, M., "Actions sketch: A novel action representation." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 1, pp. 984-989, June 2005.