A Framework for Human Activity Recognition using Pose Feature for Video Surveillance System

Alok Kumar Singh Kushwaha
Department of Computer Engineering and Application
GLA University, Mathura

Rajeev Srivastava
Department of Computer Sc. & Engineering
Indian Institute of Technology (BHU), Varanasi

ABSTRACT
In this paper, a system framework has been presented to recognize a human activity recognition approach. The proposed framework is composed of three consecutive modules: (i) detecting and locating people by background subtraction, (ii) scale invariant contour-based pose features from silhouettes (iii) finally classifying activities of people by Multiclass Support vector machine (SVM) classifier. The proposed method use approximate median filter based background–foreground separation technique to extract motion information and generate object silhouettes to activity of humans present in a scene monitored by a camera. Experimental results demonstrate that the proposed method can recognize these activities accurately for standard KTH database.

Keywords
Video Surveillance, Support Vector Machine, Approximate Median Filter

1. INTRODUCTION
The field of computer vision is continually evolving and new application areas are emerging day by day. Visual surveillance is one of the most potent and sought after application areas of computer vision and can be used for crowd flux analysis, traffic monitoring, access control to special areas, human activity recognition, anomaly detection, human computer interaction etc. Recognition of human Activity by vision still remains a challenging task before the research community. Many techniques have been proposed in past for human activity recognition. Some of the challenges faced by these technique are; automatic operation, real-time processing, changing scene background, occlusion of objects, change in illumination etc. This is also true that a technique developed to address

The challenges of one application domain may fail to address the challenges of other. So a combination of good techniques may improve the efficiency and accuracy of a system developed for a particular application domain. Many visual surveillance systems have been developed [1-2] and various surveys and frameworks on visual surveillance can be found in literature [3-5]. Recognition of human actions and activities provide important cue for human behavior analysis techniques. In sensor based activity recognition methods some smart sensory device is used to capture various activity signals for activity recognition. Vision based activity recognition methods use the spatial or temporal structure of an activity in order to recognize it. A recent survey on vision-based action representation and recognition methods can be found in [5]. Machine learning based and template based methods are popular vision based approaches for human activity recognition in videos. The machine learning based approaches for activity recognition generally solve the problem of activity recognition as a classification problem and classify an activity into one of known activity classes. For training such classifiers a number of feature types [6] and methods are used, but the drawbacks of machine learning based methods are the long training time, slow operation, constrained accuracy and it is difficult to include a new activity as well. Template based methods are good options for activity recognition in video and can be easily used because of their simplicity and robustness. Template Matching based approaches are very important among them because of their simplicity and robustness. Weinland et al. [5] provided a good survey of different human activity recognition techniques. The template matching based techniques can broadly be classified into three categories: body template based methods, feature template based methods and image template based methods. Body template based methods [6, 7] represent the spatial structure of activities with respect to the human body. In each frame of the observed video sequence, the posture of a human body is reconstructed from a variety of available image features. The action recognition is performed based on these posture estimations. This is an intuitive and biologically-plausible approach for activity recognition and supported by psychophysical work on visual interpretation of biological motion. However, in body model based representations, the resulting interest regions are linked to certain body-parts or even image coordinates. This imposes certain restrictions on recognition of different activities. In feature template based methods [8, 9], activities are recognized based on the statistics of sparse features in the image. It is a local representation of activities. It decomposes the image/video into smaller interest regions and describes each region as a separate feature. An immediate advantage of these approaches is that they neither rely on explicit body part labeling, nor on explicit human detection and localization. The approach proposed by Bobick et al. [10], used motion templates for recognizing the activities in a specific environment of aerobic exercise. They used MEI for obtaining segmented foreground and MHI for obtaining motion information in a view-specific environment. It does not give good activity recognition accuracy in outdoor environment. Moreover, their technique is capable of only identifying one activity in the scene with one actor at a time. Our present work is an extension of the work of Bobick et al. [10]. The proposed method presents templates based activity recognition. This approach considers the shape information along with the motion history for performing an activity. For obtaining the good foreground segmentation a robust approximation median filter based model is constructed. The technique can recognize the static activities like standing and sleeping as well as dynamic activities like walking, jogging, etc. in the proposed approach, covariance based matching is applied to recognize static activities and moment invariants [10, 11] are used to recognize dynamic activities. This

technique can recognize the activities of no motion such as standing and sleeping, along with those with motion such as walking, jumping.

The rest of the paper is organized as follows: section 2 gives the detailed methodology of the proposed technique used for human activity recognition, section 3 shows various experimental results and section 4 gives the conclusions.

2. THE PROPOSED METHOD

In the proposed method a template based vision activity recognition method is used. The basic purpose of this method is to accurately recognize the activities of an individual in video. For background subtraction, we use approximation median filter based approach. This model is used to subtract the background from the video frame in order to obtain the foreground. Using scale invariant contour-based pose features for different activities are created. (iii) finally classifying activities of people by Multiclass Support vector machine (SVM) classifier. The proposed technique has three basic steps as given below:

(1) Moving object segmentation.

(2) Feature Extraction

(3) Classifier

2.1 Moving Object Segmentation

In the proposed technique, we use approximation median filter based approach for background subtraction. The major advantages of the approximation median filter based method [13] are (i) its computational efficiency (ii) its robustness to external noise (iii) easier implementation (iv) and automatic detection of new appearances. The basic step of approximation median filter based approach is as follow:

Step I: Frame Differencing

For background subtraction the frame difference $FD_n(i,j)$ is obtained by taken the absolute difference two consecutive frames (n-1) & n. If the frame difference, $FD_n(i,j)$ value is less than some chosen threshold value, then $FD_n(i,j)$ is set to zero value. This process can be written as follow:-

For every pixel location (i, j) ϵ the co-ordinate of frame

$$FD_n(i,j) = |f_n(i,j) - f_{n-1}(i,j)| \quad (1)$$
$$If\ FD_n(i,j) < V_{thr}$$
$$FD_n(i,j) = 0$$

Step II: Background Modeling

In this step, check the background modeling condition as follows:

$$If\ (f_n(i,j) > f_{n-1}(i,j)) \quad (2)$$
$$f_{n-1}(i,j) = f_{n-1}(i,j) + 1$$
$$else\ f_{n-1}(i,j) = f_{n-1}(i,j) - 1$$

Here, $f_n(i,j)$ is the value of (i, j)th pixel of nth frame and $f_{n-1}(i,j)$ is the value of (i, j)th pixel of (n-1)th frame, V_{thr} is a threshold value and $FD_n(i,j)$ is the frames difference.

2.2 Features Extraction

Use the distance signal feature for activities representation and classification. The foreground image sequence (which is obtained in section 2.1) is used to extract the distance signal feature.

2.2.1 Distance Signal Feature

In this section, find out the distance signal feature using contour points of the silhouette for different key poses (sitting, standing, sleeping etc.). A binary silhouette is obtained in section 3.1 by human silhouette extraction techniques, e.g. background subtraction. We have chosen Dedeog˘ lu et al. [14] approach for contour-based distance signal feature extraction (see Fig. 1) for different key poses (sitting, sleeping etc.), which is described briefly in the following.

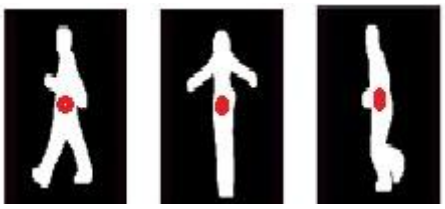


Fig. 1: Sequence of key poses of several activities (walking, jogging, running) to obtain contour based distance feature in some selected frames (KTHDB) [12]

At first, the contour points A= {a1; a2… an} of the silhouette need to be obtained. For this purpose, contour extraction is applied on the border using Suzuki et al. [15] approach.

Secondly, the centre of mass Cm = ($\bar{x}$, $\bar{y}$) of the silhouette's contour points is calculated with respect to the n number of points:

$$\bar{x} = \frac{\sum_{w=1}^{n} x_w}{n}, \quad \bar{y} = \frac{\sum_{w=1}^{n} y_w}{n} \quad (3)$$

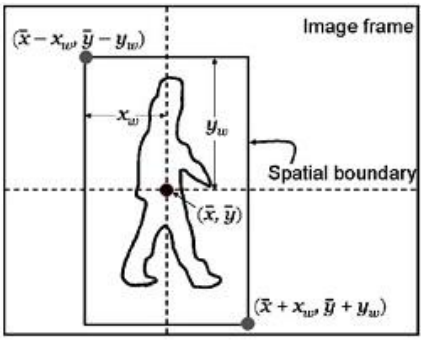


Fig. 2: Activity boundary definitions

Thirdly, the distance signal D= {d1, d2, d3……dn} is generated by determining the Euclidean distance between each contour point and the centre of mass (see fig 2.). Contour points should be considered always in the same order. For instance, the set of points can start at the most left point with equal y-axis value as the centre of mass, and follow a clockwise order.

$$d_i = \|C_m - a_i\|, \quad \forall i \in [1.....n] \quad (4)$$

Finally, scale-invariance is obtained by fixing the size of the distance signal D, sub-sampling the feature size to a constant length L and normalizing its values to unit sum.

$$\overline{D}[i] = D\left\lceil i * \frac{n}{L} \right\rceil, \quad \forall i \in [1.....L] \qquad (5)$$

$$\overline{\overline{D}}[i] = \frac{\overline{D}[i]}{\sum_{i=1}^{L} \overline{D}[i]}, \quad \forall i \in [1.....L] \qquad (6)$$

2.3 Classifier Training And Testing

After having computed features from a video, the classifier is trained and tested with these video. To model and classify activities, multi-class SVM classifiers have been used. First training of classifier is performed. The obtained training labels are supplied into the classifier then after testing is performed. The activities in the testing video are performed with the help of training labels. Finally, different test labels have been obtained for the test video of human activities. Consider the pattern recognition problem of training samples (x1, y1), (x2, y2),…..,(xl, yl), where xi, i = 1, 2,….,l is a vector and yi ∈ {1, 2…….,k} represents the class of samples. The multi-class support vector machines (SVM) [16] require the solution of the following optimization problem: minimize

$$\phi(\omega, \xi) = \frac{1}{2} \sum_{m=1}^{k} \omega_m . \omega_m + C \sum_{i=1}^{l} \sum_{m \neq y} \xi_i^m \qquad (7)$$

with constraints

$$(\omega_{y_i} . x_i) + b_{y_i} \geq (\omega_m . x_i) + b_m + 2 - \xi_i^m, \quad \xi_i^m \geq 0,$$
$$i = \{1, 2.....,l\}, m \in \{1, 2, .., k\} \setminus y_i \qquad (8)$$

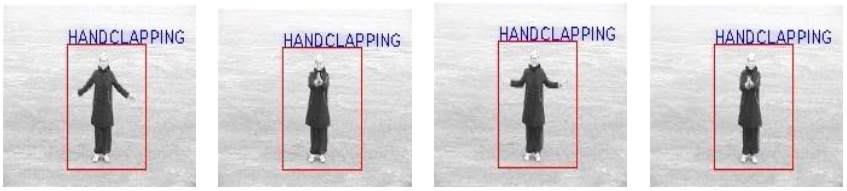
where C is the penalty parameter, l is the number of training data, k is the number of classes, yi is the class of the ith training data ω points perpendicular to the separating hyper plane, b is the offset parameter to increase the margin, and ξ is the degree of misclassification of the datum xi. This gives the decision function:

$$f(x) = \arg \max_{m=1,...,k} [\omega_m . x + b_m] \qquad (9)$$

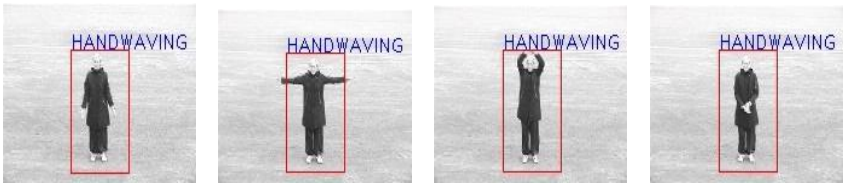
3 EXPERIMENTAL RESULTS

C++ and OpenCv libraries are used to implement the proposed human activity recognition method on a computer with 4 GB RAM and 2.53 GHhz Core i3 processor. The proposed method for activity recognition have been tested with standard KTH database [12] and WVU multi-view [17] dataset.

In Fig.3, we have shown activity recognition with standard KTH database [12]. This database includes two activities like handclapping, hand-waving. For this database also, the proposed method performs well. Moreover, this database not only contains activities involving leg motion (like jogging, running and walking) but it also contains activities involving hand motion (like boxing, handclapping and hand-waving).



(a)



(b)

Figure 3. Recognition of Activities with KTH database [12] (a) Hand-clapping (b) Hand-waving.

Table 1. Confusion Matrix of the proposed method for activity recognition considering 50 instances of each activity [12]

Recognized Instances → / **Total Instances** ↓	**Boxing**	**HandClapping**	**HandWaving**	**Jogging**	**Running**	**Walking**
Boxing (50)	46	2	1	1	0	0
HandClapping (50)	2	42	2	2	1	1
HandWaving (50)	0	0	48	2	0	0
Jogging (50)	1	1	0	46	1	1
Running (50)	0	0	1	5	43	1
Walking (50)	0	0	0	2	2	46

Table 2. Confusion Matrix of the proposed method for activity recognition considering 50 instances of each activity [17]

Recognized Instances → / **Total Instances ↓**	**Walking**	**Sitting**	**Sleeping**	**Standing**	**Bending**
Walking (50)	44	0	0	1	0
Sitting (50)	0	48	0	1	1
Sleeping (50)	0	0	43	6	1
Standing (50)	0	0	4	43	0
Bending (50)	0	2	1	1	46

Table1. Presents confusion matrix for the proposed method in performing different activities. Total 5 different activities have been considered. Take 50 instances of each activity for experiment and confusion matrix shows that how these instances are recognized. The average recognition accuracy of the proposed method is revealed as 88.52%. Similar way, table 2 show the confusion matrix of the proposed method for WVU multi-view [17] dataset. The recognition rate of the proposed method is 91.6%

4 CONCLUSIONS

We propose a human activity recognition technique based on three consecutive modules: (i) detecting and locating people by background subtraction, (ii) scale invariant contour-based pose features from silhouettes (iii) finally classifying activities of people by Multiclass Support vector machine (SVM) classifier. The experimental results demonstrate that the proposed method: (i) accurately recognizes different activities in various video frames, (ii) is suitable for static activities (like sitting, sleeping, standing, bending) as well as for dynamic activities (like jogging, walking), (iii) is pose invariant, frontal view is not necessary, (iv) can recognize activities in real outdoor and indoor environment both, (v) is suitable for operation in outdoor environment in the presence of shadow, (iv) is suitable for activities not only involving leg motion (like jogging, running, walking) but also for activities involving hand motion (like boxing, hand-clapping, hand-waving).We observed activity recognition accuracy as 88.52%.Since the observed execution speed is 25 frames/sec, the proposed technique is suitable for real time video surveillance and other monitoring application.

5 REFERENCES

[1] Enficiaud, R., Lienard, B., Allezard, N., Sebbe Serge Beucher, R., Desurmont, X., Sayd, P., and Delaigle, J., 2006. CLOVIS - A generic framework for general purpose visual surveillance applications.

[2] Chen, P.Y., Lin, H. M., Chen, W. T., and Tseng, Y. C., 2010. Demo abstract: a multi-view visual surveillance system based on angle coverage, in Proc.in the 8th ACM Conference on Embedded Networked Sensor System.

[3] Valera, M., and Velastin, S.A., 2005. Intelligent distributed surveillance systems: a review, Int. j. Vision, Image and Signal Processing, 152(2): 192-204.

[4] Srinivasan, K., Porkumaran, K. and Sainarayanan, G., 2009. Intelligent human body tracking, modeling, and activity analysis of video surveillance system: A survey, Int. J. of Convergence in Engineering, Technology and Science, 1: 1-8.

[5] Weinland, D. and Ronfard, R., 2011. A survey of vision based methods for action representation, segmentation, and recognition,Computer Vision and Image Understanding, 115(2): 529–551.

[6] Junejo, I., Dexter, E., Laptev, I. and Perez, P. View-independent action recognition from temporal self-similarities, IEEE Trans. On Pattern Analysis and Machine Intelligence, in press.

[7] Laptev, I., Caputo, B., Schuldt, C., and Lindeberg, T., 2007. Local velocity adapted motion events for spatio-temporal recognition, 108: 207-229.

[8] Ke, Y., Sukthankar, R., and Hebert, M., 2010. Volumetric features for video event detection, Int. J. of Computer Vision.

[9] Technical Report CMU-CS-08-113., 2008. Volumetric features for video event detection.

[10] Bobick, A.F., and Davis, J.W., 2001. The recognition of human movement using temporal templates, IEEE Trans. Pattern Analysis and Machine Intelligence, 23(3): 257-267.

[11] Hu, M., 1962. Visual pattern recognition by moment invariants, IRE Trans. Information Theory, 8(2): 179-187.

[12] KTH Research Project Activity Database. Available at: http://www.nada.kth.se/cvap/actions

[13] McFarlane, N., and Schofield, C., 1995. Segmentation and tracking of piglets in images, Machine Vision Application, 8(3): 187-193.

[14] Y. Dedeog˘ lu, B. Töreyin, U. Güdükbay, A. Çetin, "Silhouette-based method for object classification and human action recognition in video", Computer Vision in Human-Computer Interaction, Lecture Notes in Computer Science, 3979. Springer, Berlin/Heidelberg, pp. 64–77, 2006.

[15] S. Suzuki, K. be, "Topological structural analysis of digitized binary images by border following", Comput. Vision Graphics Image Process, Vol. 30, pp. 32–46, 1985.

[16] J. Westons and C. Wtkins, "Support Vector Machines for Multiclass Pattern Recognition," Proc. 7th European Symposium on Artificial Neural Networks, pp. 219-224, 1999.

[17] V. Kulathumani, WVU Multi-View Action Recognition DatasetAvailable on: http://csee.wvu.edu/~vkkulathumani/wvu-action.html#download2