Automatic Activity Recognition for Video Surveillance

J. Arunnehru
Dept. of Computer Science and Engg.,
Annamalai University

M. Kalaiselvi Geetha
Dept. of Computer Science and Engg.,
Annamalai University

ABSTRACT

Activity recognition is having a wide range of applications in automated surveillance and is an active research topic among computer vision community. In this paper, an activity recognition approach is proposed. Motion information is extracted from the difference image based on Region of Interest (ROI) using 18-Dimensional features called Block Intensity Vector (BIV). The experiments are carried out on the KTH dataset considering four activities viz., (walking, running, waving and boxing) with SVM. The approach shows an overall performance of 94.58% in recognizing the actions performed. Experimental results show that the proposed approach is comparable with the existing methods.

Keywords:

Video Surveillance, Activity Recognition, Gesture Recognition, Support Vector Machines, Difference Image

1. INTRODUCTION

Video surveillance is attracting much of the researchers attention due to its wide application in human computer interaction, security, unusual activity recognition etc., The most fascinating area of research in the fields of artificial intelligence and pattern recognition is to understand the human activities automatically from video sequences. It is a complex process to recognize human action due to many factors such as variation in speed, postures and clothing and their appearance is affected by scene factors like occlusion or illumination [1]. It is clear that developing a good algorithm to solve action recognition problem that gives huge potential becomes essential for a large number of applications like video surveillance, gesture recognition, human-computer interaction, robot learning and control, etc., Many approaches for learning and recognizing human actions from image measurements are seen in [2, 3, 4]. This paper deals with activity recognition that aims to understand human activities from video sequences.

1.1 Related Work

An independent approach is used to identify the different actions by 3-D descriptors and 2-D binary shapes are captured by multiple cameras with SVM [5]. Appearance based as low level descriptors are extracted from interest points and combined with global descriptors for robust classification and recognition of human actions from movies [6]. Temporal structure and local features are combined to recognize the complex events at different temporal scales using a bag of word classifier [7]. The human activity recognition in video using background modeling based technique, motion history frames are used to construct the object shape information and spatial-temporal template for different human activities like walking, running, bending, sleeping and jumping in [8]. An optical flow and spatio-temporal gradient features are used in appearance based approach and geometric features are used in pose based approach is compared in [9] and shows that is considerably better than appearance based approach. The 3D joint trajectories with histogram are used to recognize the human activities [10]. In bag of words approach, scene flow features (depth and optical flow) instead of HoG/HoF features are used to recognize the human activity on stereo images in clutter background [11].

1.2 Outline of the work

This paper deals with activity recognition for video surveillance. The proposed method is evaluated using the KTH action dataset with the persons showing actions such as walking, running, waving and boxing. Difference image is extracted from two consecutive frames. Region of Interest (ROI) is extracted by dividing the difference image into three blocks. Block Intensity Vector (BIV) is extracted from ROIs and the feature vector is fed to the SVM classifier for recognizing the activity performed. A detailed description of the feature extraction procedure is explained in the following section.

The rest of the paper is organized as follows. Section 2 describes the features that are intrinsic in the classification process and the related discussions. Section 3 furnishes a brief introduction on the SVM classifier. The proposed algorithm is explained in Section 4. Experimental results are presented in Section 5 and finally Section 6 concludes the paper.

2. FEATURE EXTRACTION

Video sequence is a rich versatile information. Appropriate recognition and representation of this information is needed for further processing. For that purpose, a feature is a descriptive aspect extracted from a video stream. Recognizing or classifying an appropriate activity performed by an object in a video sequence relies heavily on competent use of these features that provide discriminative information useful for high level recognition. The following subsections present the description of the feature used in this work.

2.1 Frame Differencing

Frame differencing is defined by the differences between successive frames in time, as an alternative of subtracting a predefined background on-the fly, the frame subtraction method considers every pair of frames of time t and $t+1$

to extract any motion information in it. In order to locate the regions of interest, by simply subtracting the current frame with previous frame on a pixel by pixel basis, Fig. 1 (a), Fig. 1 (b) shows the two consecutive frames of the KTH dataset. The difference image of the motion information is shown in Fig. 1 (c). Then the absolute value of the difference image is compared with a predetermined threshold value. The difference image at time t is given by:

$$\begin{gathered} D_t(i,j) = |I_t(i,j) - I_{t+1}(i,j)| \\ 1 \leq i \leq w, 1 \leq j \leq h \end{gathered} \tag{1}$$

$I_k(i,j)$ is the intensity of the pixel (i,j) in the k^{th} frame, w and h are the width and height of the image respectively. The proposed approach uses an image size of 160 x 120. The region of motion information obtained is considered as the Region of Interest.

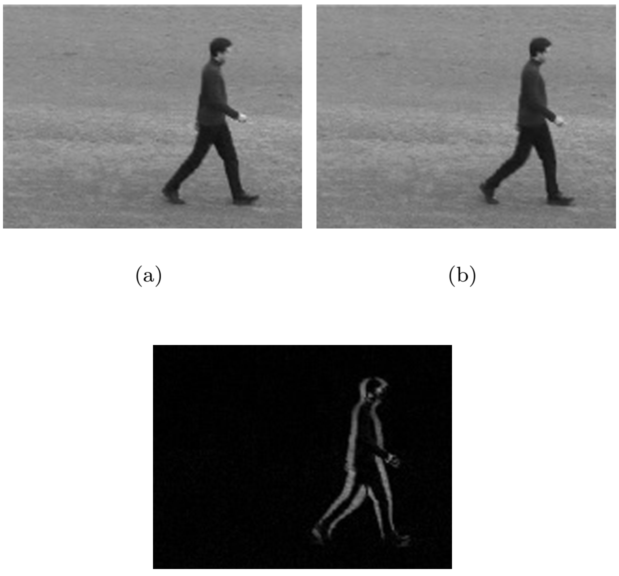


Fig. 1. (a), (b) Two Consecutive frames. (c) Difference image of (a) and (b) from KTH dataset.

2.2 Motion Feature

Motion information T_k or difference image is calculated using:

$$T_k(i,j) = \begin{cases} 1, & \text{if } D_k(i,j) > t; \\ 0, & \text{Otherwise;} \end{cases} \tag{2}$$

where t is the threshold. The value of $t = 30$ has been used in the experiments. To capture the dynamic information, motion is extracted from the difference image D_t as in Eq. 1. Motion information obtained for walking and running are shown in Fig. 2 and Fig. 3 respectively.

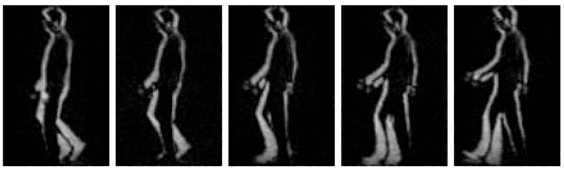


Fig. 2. Motion information extracted for walking.

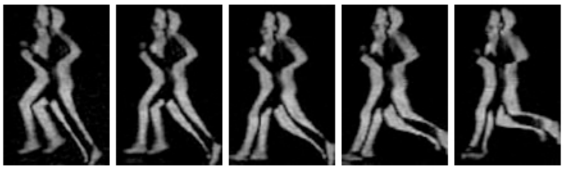


Fig. 3. Motion information extracted for running.

2.3 Block Intensity Vector (BIV)

The proposed Block Intensity Vector for action recognition is described here Fig. 4 (a) shows the ROI identified and Fig. 4 (b) is the extracted ROI. Fig. 4 (c) is the block description of the extracted ROI. The ROI is identified from the difference image calculated.

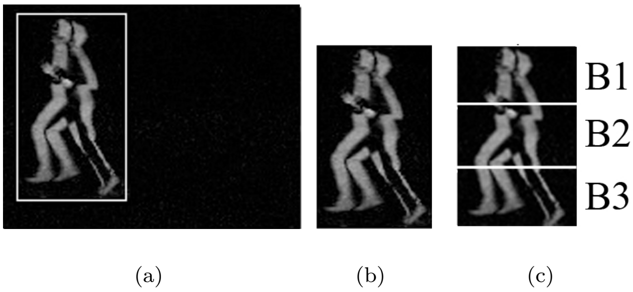
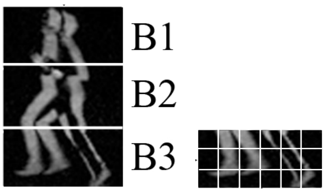


Fig. 4. (a) Motion information. (b) ROI extracted from (a). (c) Block description of (b).

Since the aim of this work is activity recognition, it is seen that human perform actions mainly with their arms and legs. So, to minimize computation, this work divides the ROIs into three blocks B1, B2 and B3 comprising head, torso and leg regions as in Fig. 4 (c) and further processing continues only in the regions that exhibits maximum motion. To identify the block with maximum motion, the intensity values of pixels in each block are calculated. The block with maximum intensity is considered as the region with heavy motion.



(a) (b)

Fig. 5. (a) Three block representation of ROI. (b) 3 x 6 representation of (a).

It is seen that maximum intensity values are seen in B3 region as seen in Fig. 5 (a). So, B3 region only is considered for further processing. B3 region is again divided into 3 x 6 blocks as in Fig. 5 (b) and the intensity value of pixels in each block is calculated to form 18 dimensional feature vector called Block Intensity Vector (BIV).

3. SUPPORT VECTOR MACHINE

Support Vector Machines (SVMs) are useful methods for data classification. Which have newly gained popularity within visual pattern recognition [12, 13]. In this section, SVM theory is discussed briefly. A classification task typically involves separating data into training and testing sets. In the each occurrence in the training set contains one class labels and several observed variables. The aim of SVM is to construct a model based on the training data to predict the target values of the test. $(x_i, y_i), i = 1, ..., l$ where $x_i \in \Re^n$ is a feature vector and $y_i \in \{+1, -1\}$ is a class label.

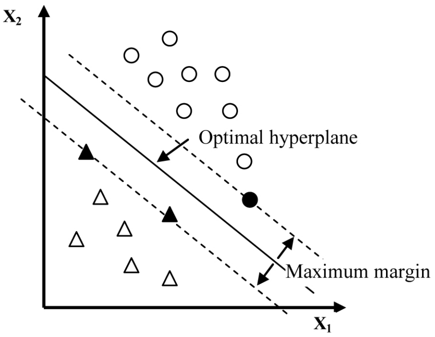


Fig. 6. Representation of Hyperplane.

The goal of SVM is to separate the data with the hyperplane using a kernel trick [14, 15]. The Distance between the nearest points on hyperplane to origin can be found by maximizing the x on the hyperplane. The hyperplane $w.x + b = 0$ is used to separate the two classes in feature space, where w is the coefficient of vector and b is the scalar. Such that $y_i(w.x_i+b) \geq 1$, $\forall i$. The maximum margin between two classes is given by $M = 2/||w||$ as shown in Fig. 6. To solve the minimization problem by using Lagrange multipliers $\alpha_i (i = 1, ...m)$ to obtain the optimal values for w and b is given in Eq. 3.

$$f(x) = sgn\left(\sum_{i=1}^{m} \alpha_i y_i \; K(x_i, x) + b\right) \tag{3}$$

where α_i and b are discussed in SVC learning algorithm [16], x_i is a nonzero value, α_i are the support vectors, $K(x, y) = x.y$ is used to make the optimal separating hyperplane in input space $\Re^n$. Eq. 4 and Eq. 5 are used to refine the hyperplane in minimization problem.

$$\min_{w,b,\xi} \; \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{4}$$

$$y_i(w^T \phi(x_i) + b \geq 1 - \xi_i, \xi_i \geq 0 \tag{5}$$

3.1 Kernel Trick

If data is linear, a separating hyper plane may be used to divide the data. In some case the data are distant from linear and the datasets are inseparable. To permit for these kernels are used for non-linear map the input data to a high-dimensional space. This is done by a kernel function, and it has its own set of parameters, by translating this back to the original feature space as shown in the Fig. 7. *Feature Space:* Transforming the data into a feature space makes it feasible to define a resemblance measure on the basis of the dot product. If the feature space is selected suitably, pattern recognition can be easy [17]. Here training vectors x_i are mapped into a higher dimensional space by the function φ.

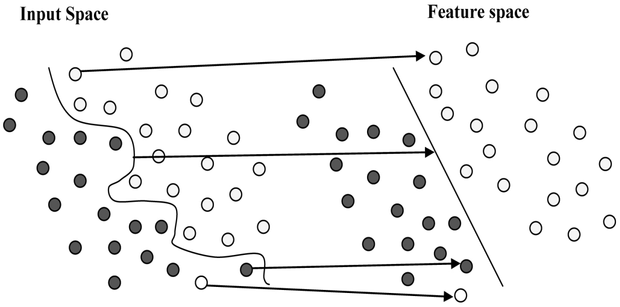


Fig. 7. Input space to Feature space.

3.2 Kernel Functions

The idea of the kernel function is to permit operations to be performed in the input space relatively than the potentially high dimensional feature space. The inner product does not need to be evaluated in the feature space. Kernel functions are

Linear:

$$K(x_i, x_j) = x^T{}_i x_j \tag{6}$$

Polynomial:

$$K(x_i, x_j) = (\gamma x^T{}_i x_j + \gamma)^d, \gamma > 0 \tag{7}$$

Radial Basis Function (RBF):

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2), \gamma > 0 \tag{8}$$

Sigmoid:

$$K(x_i, x_j) = \tanh(\gamma {x^T}_i - x_j + r) \tag{9}$$

Here, γ, r, and d are kernel parameters. In multi-class SVM, several classifier is used for training and combining their results. There are many strategies for combining SVM, two common methods are, 1) one per class and 2) pairwise coupling. The pairwise coupling method trains k - is a number of classes. $k(k-1)/2$, when the number of the classes is high the pairwise coupling method requires the training of a huge number of SVM, for classification, by using a multi-class SVM, to construct according to one per class strategy to train the classifier.

4. PROPOSED APPROACH

The workflow of the proposed approach is seen in Fig. 8. The approach used KTH dataset. The persons showing actions such as walking, running, waving and boxing are considered for analysis. The video sequence is converted into frames in .jpg format. To beginwith the first frame is compared with the successive frames to compute frame differencing as discussed in section 2.3. This gives the foreground image that exhibit motion. This region is considered as Region of Interest (ROI). The ROI is divided into 3 x 1 blocks each of size 30 x 60 pixels. A BIV is extracted from the selected block, which have maximum intensity value and is divided into 3 x 6 subblocks each of size 10 x 10 pixels. The proposed algorithm for action recognition is discussed here.

4.1 Algorithm

(1) The difference image is calculated from consecutive frames in video sequence as in Eq. 1.

(2) Motion information is extracted from the difference image calculated with Eq. 2.

(3) BIV is extracted as explained in section 2.3.

(4) The feature vector is fed to SVM for further analysis.

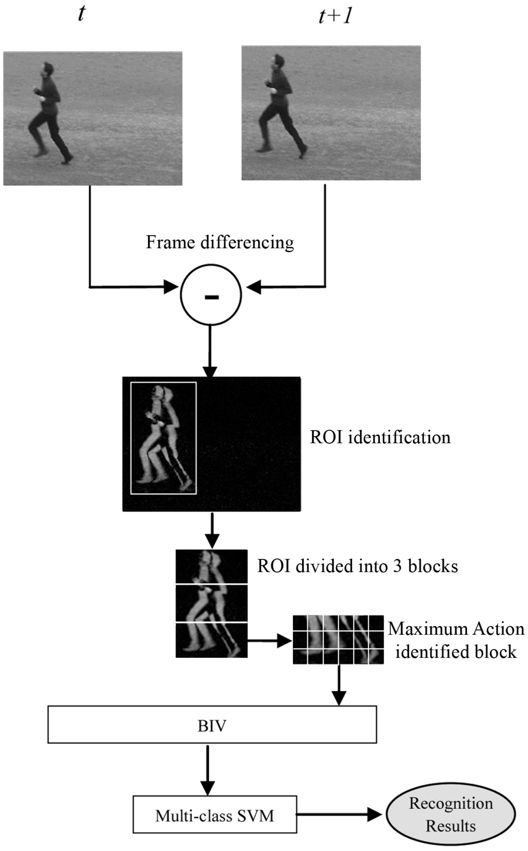


Fig. 8. Workflow of the proposed approach.

5. EXPERIMENTAL RESULTS

The experiments are carried out in C++ with OpenCV 2.2 in Ubuntu 12.04 operating system on a computer with Intel Xeon X3430 Processor 2.40 GHz with 4 GB RAM. The extracted BIV features are fed to LIBSVM [18] for training. Polynomial and RBF kernels are used for experimental purpose.

5.1 Dataset

From the KTH dataset, four different actions viz., walking, running, waving and boxing are considered. 25 different persons performed the actions in various scenarios like $s1$: outdoor, $s2$: outdoor with scale variations, $s3$: outdoor with different clothes and $s4$: indoors. Sample frames of action sequence are shown in the Fig. 9. The sequences were taken over homogeneous backgrounds with a static camera. Video data are at 25 fps, each video clip contains one actor performing one action in four different scenarios as explained earlier. The dataset shown in Table 1 is used for experimental purpose. Four actions walking, running, waving and boxing taken from 4 different scenarios are used in the experiments. Each clip is of 1 sec duration and for each action, a total of 40 clips are utilized, that includes all four scenarios considered in this work.

Table 1. KTH dataset with different scenarios in our proposed method

Actions	No. of clips			
	s1	s2	s3	s4
Walking	12	8	7	13
Running	11	9	8	12
Waving	13	8	7	12
Boxing	12	7	8	13

In this work, 10 persons are taken randomly from four scenarios for evaluation. The samples are divided into a training set of (7 persons), and testing set of (3 persons).

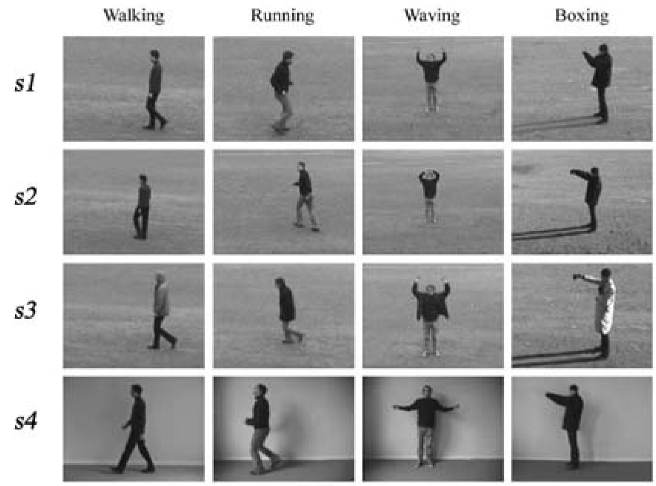


Fig. 9. Action samples from KTH dataset.

5.2 Preprocessing

For the successful feature extraction and classification, it is essential to preprocess all video sequences to remove noise and incomplete data for fine features generation. To remove noise and weaken image distortion, all frames of each action sequences are smoothed by Gaussian convolution with a kernel of size 3 x 3 and variance $\sigma = 0.5$.

5.3 Results obtained with SVM

The recognition results obtained by the proposed method on KTH dataset with Polynomial SVM and RBF SVM are summarized in a confusion matrix in Table 2 and 3, where correct responses define the main diagonal, the majority of actions are correctly classified, An average recognition rate of Polynomial kernel is **87.92%** and RBF kernel is **94.58%**. Thus the RBF kernel is showing better performance when compared to polynomial kernel.

Table 2. Confusion Matrix for POLYNOMIAL KERNEL (87.92)

	Walking	Running	Waving	Boxing
Walking	**0.92**	0.08	0.00	0.00
Running	0.15	**0.78**	0.03	0.03
Waving	0.02	0.03	**0.90**	0.05
Boxing	0.00	0.00	0.08	**0.92**

Table 3. Confusion Matrix for RBF KERNEL (94.58)

	Walking	Running	Waving	Boxing
Walking	**0.97**	0.03	0.00	0.00
Running	0.12	**0.88**	0.00	0.00
Waving	0.00	0.02	**0.97**	0.01
Boxing	0.00	0.00	0.03	**0.97**

5.4 Comparative Study

Table 4. Comparsion with various methods

Method	Accuracy(%)
Proposed approach	**94.58**
Jhuang et al. [19]	91.70
Nowozin et al. [20]	87.04
Dollar et al. [21]	81.17
Schuldt et al. [22]	71.72

To demonstrate the efficiency of the proposed approach, a comparison is made with the approaches as listed in Table 4. Based on the comparison, it is seen that the proposed method shows good results.

6. CONCLUSION AND FUTURE WORK

This paper presents an approach for activity recognition for video surveillance using Block Intensity Vector (BIV) as features, Experiments are conducted on four different actions viz., walking, running, waving and boxing from KTH action dataset. The extracted ROI from the difference image are used for classification based on motion features. The approach then evaluates the performance of BIV features in the video sequence using SVM with polynomial and RBF kernels. The system gives a good classification accuracy of 94.58% by RBF having the parameter of $C = 32$ and $\gamma = 2.0$. It is observed from the experiments that running and walking results, the system could not distinguish running and walking with high accuracy and is of future interest.

Acknowledgments

This work is being supported by **University Grants Commission of India** [F. No. 41-636/2012 (SR)].

7. REFERENCES

[1] Sadek, S., Al-Hamadi, A., Michaelis, B. and Sayed, U. 2011. Human action Recognition: A novel scheme using fuzzy log-polar histogram and temporal self-similarity, *EURASIP Journal on Advances in Signal Processing.*

[2] Poppe, R. June, 2010. A survey on vision-based human action recognition, *In Proceedings of Image and Vision Computing*, Vol. 28, pp. 976–990.

[3] Weinland, D., Ronfard, R. and Boyer, E. 2011. A survey of vision-based methods for action representation, segmentation and recognition, *In Proceedings of Computer Vision and Image Understanding* , Vol. 115, pp. 224–241.

[4] Rautaray, S. and Agrawal, A. 2012. Vision based hand gesture recognition for human computer interaction: A survey, *Artificial Intelligence Review.*

[5] Cohen, I. and Li, H. 2003. Inference of Human Postures by Classfication of 3D Human Body Shapes, *IEEE international Workshop on Analysis and Modeling of Faces and Gestures*, pp. 74–81.

[6] Ivan Laptev, Marcin Marszalek, Cordelia Schmid and Benjamin Rozenfeld. June, 2008. Learning realistic human actions from movies, *In Conference on Computer Vision and Pattern Recognition.*

[7] Juan Carlos Niebles, Chih-Wei Chen and Fei-Fei, L. 2010. Modeling temporal structure of decomposable motion segments for activity classification, *In Proceedings of the 11th European Conference of Computer Vision (ECCV), Greece.*

[8] Chandra mani Sharma, Alok Kr, Singh Kushwaha, Swati Nigam and Ashish Khare. 2011. Automatic human activity recognition in video using background modeling and spatio-temporal template matching based technique, *Proceedings of the International Conference on Advances in Computing and Artificial Intelligence*, pp. 99–101.

[9] Yao, A., Gall, J., Fanelli, G. and Van Gool, L. 2011. Does Human Action Recognition Benefit from Pose Estimation?, *In Proceedings of BMVC*, pp. 1–67.

[10] Xia, L., Chen, C. C. and Aggarwal, J. K. June, 2012. View invariant human action recognition using histograms of 3D joints, *In Proceedings of CVPR workshop on Human Activity Understanding from 3D Data (HAU3D).*

[11] Jordi Sanchez-Riera, Jan Cech and Radu Horaud. October, 2012. Action Recognition Robust to Background Clutter by Using Stereo Vision, *4th International Workshop on Video Event Categorization, Tagging and Retrieval.*

[12] Wallraven, C., Caputo, B. and Graf, A. 2003. Recognition with local features: Kernel receipe, *In Proceedings of ICCV*, pp. 257–246.

[13] Wolf, L. and Shashua, A. 2003. Kernel Principal angles for classification machines with applications to image sequences interpretation, *In Proceedings of CVPR*, pp. 635–640.

[14] Nello Cristianini and John Shawe-Taylor. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, *Cambridge University Press.*

[15] Tom Mitchell. 1997. Machine Learning, *McGraw-Hill Computer science series.*

[16] Vapnik, V. 1998. Statistical Learning Theory, *Wiley, NY.*

[17] Lewis, J. P. 2004. Tutorial on SVM, *CGIT Lab, USC.*

[18] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 1–27.

[19] Jhuang, H., Serre, T., Wolf, L. and Poggio, T. 2007. A biologically inspired system for action recognition, *In Proceedings of ICCV.*

[20] Nowozin, S., Bakir, G. and Tsuda, K. 2007. Discriminative subsequence mining for action classification, *In Proceedings of ICCV.*

[21] Dollar, P., Rabaud, V., Cottrell, G. and Belongie, S. 2005. Behavior recognition via sparse spatio-temporal features, *Proceedings of the 14thInternational Conference on Computer Communications and Networks*, pp. 65–72.

[22] Schuldt, C., Laptev, I. and Caputo, B. June, 2004. Recognizing human actions: A local SVM approach, *Pattern Recognition, Proceedings of the 17th International Conference, ICPR 04*, Vol. 3, pp. 32–36.