Pak. J. Statist.
2016 Vol. 32(2), 109-123

A STATISTICAL MODEL FOR ANNOTATING VIDEOS WITH HUMAN ACTIONS

M.U.G. Khan[1], A. Nasir[1§], O. Riaz[2], Y. Gotoh[3] and **M. Amiruddin[1]**
[1] Deptt. of Computer Science & Engineering, Al-Khwarizmi Institute of Computer Science, University of Engineering & Technology, Lahore, Pakistan
[2] Deptt. of Computer Science, Islamia University, Bahawalpur. Pakistan
[3] Department of Computer Science, The University of Sheffield, UK
[§] Corresponding author Email: nasirbhutta1@gmail.com

ABSTRACT

This contribution addresses the approach to recognize single and multiple human actions in video streams. This work introduces a novel action recognition algorithm with normalization enhancements. Initially feature vectors are extracted using 2D SIFT features. Bag of Words model is extended with a new normalization technique on the visual vocabulary to make the dimensions same so that the actions would be easier to read and extract. This normalization technique vastly improves the results from the state of the art methods. HMM based model is developed for training and testing of six basic actions present in the KTH human action dataset. By comparing our work with previously applied models, results display that our approach vastly improves the accuracy of the existing methods of action recognition.

INTRODUCTION

Human action recognition is a widespread area of research given the advancement in digital technology. It is even more so given the online exponentially increasing trend of video data. To extract information from the videos humans need to view the videos. As humans are prone to errors our goal is to provide an enhancement algorithm to already established methods and techniques to ease the recognition of events and behaviors for retrieval, alerting and summarization of the data.

Action recognition representation can be categorized as: flow based approaches [12], spatio-temporal shape template based approaches [25], tracking based approaches [4] and interest points based approaches [51]. In flow based approaches optical flow is used for describing motion descriptors. Optical flow computation is used to detect motion descriptors. Optical flow is quite sensitive to noise and changing background so it cannot truly detect the changing action in different frames. Spatio-temporal shape template based algorithms describe the human action recognition as a 3D action frame and features are extracted using 3D volume approaches. The spatio temporal shaped approaches are good for small videos but where a huge dataset is concerned the computational costs are too high. Tracking based approaches also has the same problems as spatio-temporal based approaches that they have a huge computational cost. Interest point based action recognition algorithms have very short feature vectors. This leads to very low computational costs; hence it is the most popular form of gesture recognition.

One of the most widespread used techniques in the field of computer vision and especially video detection and gesture recognition is using bag of visual words algorithm [57]. In this algorithm videos are treated as documents and the features and gestures in the videos are regarded as words. This approach is quite strong against change and noise. After using bag of words some classification algorithm such as HMM or SVM are used.

In this work 2DSIFT [23] is used for detecting interest points where the extracted features are invariant to scale, location and orientation changes. 2D sift limits the size of feature vectors which consumes less computational time. Bag of words features are used to build visual vocabulary and HMM classification algorithm is used. Results are tested on the famous KTH dataset.

The rest of the paper is organized as follows: the next section reviews previous related work, then the proposed system is presented followed by the experiments and results, and then the conclusion.

RELATED WORK

There are basically three types of method division on which action recognition is done:

Human model based methods employ a full 3D (or 2D) model of human body parts, and action recognition is done using information on body part positioning as well as movements. A significant amount of research [26] is devoted to action recognition using trajectories of joint positions, body parts, or landmark points on the human body with or without a prior model of human kinematics, e.g., [36][27], [46]. The localization of body parts in movies has been investigated in the past (e.g. [32], [13]) and some works have shown impressive results. However, the detection of body parts is a difficult problem in itself, and results especially for the case of realistic and less constrained video data remain limited in their applicability. Some recent approaches that are able to provide more robust results (e.g., [3], [39]), use strong prior knowledge by assuming particular motion patterns in order to improve tracking of body parts. However, this also limits their application to action recognition.

Holistic methods use knowledge about the localization of humans in video and consequently learn an action model that captures characteristic, global body movements without any notion of body parts. In general, holistic approaches can be roughly divided into two categories. The first category employs shape masks or silhouette information, stemming from background subtraction or difference images, to represent actions. The second category is mainly based on shape and optical flow information.

Several approaches for action recognition use human shape masks and silhouette information to represent the human body and its dynamics. The method discussed in [44] is among the first to propose silhouette images. Their representation computes a grid over the silhouette and computes for each cell the ratio of foreground to background pixels. The grid representations are quantized into a vocabulary, and tennis actions are then learned as sequences of\words" using hidden Markov models (HMM) [31]. [7] Uses shape masks from difference images to detect human actions.

As action representation, the authors employ so-called motion energy images (MEI) and motion history images (MHI). More precisely, MEIs are binary masks that indicate regions of motion, and MHIs weight these regions according to the point in time when they occurred (the more recent, the higher the weight). This approach is the first to introduce the idea of temporal templates for action recognition. The system mentioned in[38] detected tennis forehand strokes by matching a set of hand drawn key postures together with annotated body joint positions to edge information in a video sequence. Positions of joints are then tracked between the key frames using silhouette. An action model based on space-time shapes from silhouette information is introduced by [6], [15]. Silhouette information is computed using background subtraction. The authors use properties of the solution to the Poisson equation to extract features such as local saliency, action dynamics, shape structure and orientation. Chunks of 10 frames length are then described by a high-dimensional feature vector. During classification, these chunks are matched in a sliding window fashion to space-time shapes in test sequences.

Another work that uses space-time shapes of humans if discussed in [45]. Spatio-temporal shapes are obtained from contour information using background subtraction, similar to [6]. For a robust representation, actions are then represented by sets of characteristic points (such as saddle, valley, ridge, peak, pit points) on the surface of the shape. In order to recognize actions, the authors propose to match spatio-temporal shapes by computing homographs using point-to-point correspondences. The system mentioned in [41] introduces an order less representation for action recognition using a set of silhouette exemplars. Action sequences are represented as vectors of minimum distance between silhouettes in the set of exemplars and in the sequence. Final classification is done using Bayes classifier with Gaussians to model action classes. In addition to silhouette information, the authors also employ the Chamfer distance measure to match silhouette exemplars directly to edge information in test sequences. Foreground shape masks based on motion information in chunks of video data are employed by [47]. Silhouettes are also a popular representation for surveillance applications [16],[37]. Since cameras are in general static, background subtraction techniques can be employed to compute silhouette information. In order to cope with more challenging video data and camera motion, [33] employs a human tracker and camera motion estimation to compute shape information. However, to deal with noisy and imprecise segmentation information, a more robust classification method is used as well.

Another way to match space-time shape models to cluttered image data with heterogeneous background is demonstrated by [32]. The authors over segment video sequences using color information. Volumetric and optical flow features are then matched to action templates in form of space-time shapes. To account for occlusion and actor variability, [32] extends their template to an action part model using pictorial structures. Silhouettes provide strong cues for action recognition. Nevertheless, they are difficult to compute in the presence of clutter and camera motion. Furthermore, they only describe the outer contours of a person and thus lack discriminative power for actions that include self-occlusions. Human-centric approaches based on optical flow and generic shape information form another sub-class of holistic methods. As one of the first works in this direction, [29] propose a human tracking framework along with an action representation using spatio-temporal grids of optical flow magnitudes. The action descriptor is computed for periodic motion patterns. By matching against reference

motion templates of known periodic actions (e.g., walking, running, swimming, skiing) the final action can be determined. In another approach purely based on optical flow, [11] track soccer players in videos and compute a descriptor on the stabilized tracks using blurred optical flow. Their descriptor separates x and y flow as well as positive and negative components into four different channels. For classification, a test sequence is frame-wise aligned to a database of stored, annotated actions. Further experiments include tennis and ballet sequences as well as synthesis experiments.

The same human-centric representation based on optical flow and human tracks for action recognition is employed by [12]. As classification framework, the authors use a two-layered AdaBoost [14] variant. In the first step, intermediate features are learned by selecting discriminative pixel flow values in small spatio-temporal blocks. The final classifier is then learned from previously aggregated intermediate features. Evaluations are carried out on four datasets: KTH, Weizmann, soccer, and a ballet dataset. The system mentioned in [34] proposed an approach using flow features in a template matching framework. Spatio-temporal regularity flow information is used as feature type. Regularity flow shows improvement over optical flow since it globally minimizes the overall sum of gradients in the sequence. Rodriguez et al. learn cuboid templates by aligning training samples via correlation. For classification, test sequences are correlated with the learned template via generalized Fourier transform that allows for vectorial values. Results are demonstrated on the KTH dataset, for facial expressions, as well as on custom movie and sports actions. To localize humans performing actions such as sit down, stand up, grab cup and close laptop, [32] use a forward features selection framework and learn a classifier based on optical flow features. Spatio-temporal Haar features on optical flow components are efficiently computed using an integral video structure. During learning, a discriminative set of features are greedily chosen to optimally classify actions which are represented as spatio-temporal cuboidal regions. For classification, the authors perform a sliding window approach and classify each position as containing a particular action or not. A method purely based on shape information is presented in [24]. In their experiments, Lu and Little track soccer or ice-hockey players and represent each frame by a descriptor using histograms of oriented gradients. They then employ principal component analysis (PCA) [28] to reduce dimensionality. An HMM with a few states models actions such as running, skating, left, right etc. Hybrid representations combine optical flow with appearance information [35].Use optical flow information and Gabor later responses in a human-centric framework. For each frame, both types of information are weighted and concatenated.

Local feature methods are entirely based on descriptors of local regions in a video, no prior knowledge about human positioning or of any of its limbs is given. Feature detectors usually select characteristic spatio-temporal locations and scales in videos by maximizing specific saliency functions [20], [19] are the first to propose a feature detector based on a spatio-temporal extension of the Harris corner criterion [17]. The corner criterion is based on the eigenvalues of a spatio-temporal second-moment matrix at each video point. Local maxima indicate points of interest. The authors note the importance of using separate spatial and temporal scale values since spatial and temporal extent of events are in general independent. Results of detecting Harris interest points in an outdoor image sequence of a person walking.

In [10] authors argue that true spatio-temporal corner points (according to the Harris criterion) are relatively rare, while enough characteristic motion is still present. Therefore, they design their interest point detector to yield denser coverage in videos. Their method employs spatial Gaussian kernels. As for 3D Harris, local maxima give final interesting positions.

The Hessian 3D detector is proposed by [42] as spatio-temporal extension of the Hessian saliency measure applied for blob detection in images [5]. The authors aim at a rather dense, scale-invariant, and computationally efficient interest points to determine salient features by considering global information [43]. For this, video sequences are represented as dynamic texture with a latent representation and a dynamic generation model. This allows synthesizing motion, but also to identify important regions in motion. The dynamic model is approximated as linear transformation.

HOG and HOF descriptors are introduced by [22]. To characterize local motion and appearance, the authors combine histograms of oriented spatial gradients (HOG) and histograms of optical flow (HOF) in a late fusion approach. The histograms are accumulated in the space-time neighborhood of detected interest points. Each local region is subdivided into a N_N_M grid of cells; for each cell, 4-bin HOG histograms and a 5-bin HOF histogram are computed. The normalized cell histograms are concatenated into the final HOG and HOF descriptors. An extension of the image SIFT descriptor [23] to 3D was proposed by [43]. For a set of randomly sampled positions, spatio-temporal gradients are computed in the local neighborhood of each position. Each pixel in the neighborhood is weighted by a Gaussian centered on the given position and votes into an M _M _M grid of histograms of oriented gradients. For orientation quantization, the authors represent gradients in spherical coordinates; that are divided into a 8 by 4 histogram. The system mentioned in [42] proposed the extended SURF (ESURF) descriptor which extends the image SURF descriptor [4] to videos. Like in previous approaches, the authors divide 3D patches into a grid of local M _ M _ M histograms. Each cell is represented by a vector of weighted sums of uniformly sampled responses of Haar-wavelets along the three axes.

[2] proposed a method with four approaches for this task, (i) histograms of SIFT features as feature vectors and Bhatacharya distance for similarity detection (ii) feature vector is combination of SIFT features alone, while for matching a basic descriptor matching algorithm (iii) IR based approach using SIFT features and (iv) affine invariant SIFT features as feature vector.

METHODOLOGY

Our proposed methodology consists of five steps: Initially, interest points are detected using scale invariant feature transform (SIFT). Secondly, SIFT descriptors are deployed for description of interest points. Then, a visual dictionary is constructed using K-Means clustering algorithm. To further augment the already built dictionary, we further build a vocabulary of visual words using bag of words approach. Finally a classification model is generated using hidden Markov model based approach.

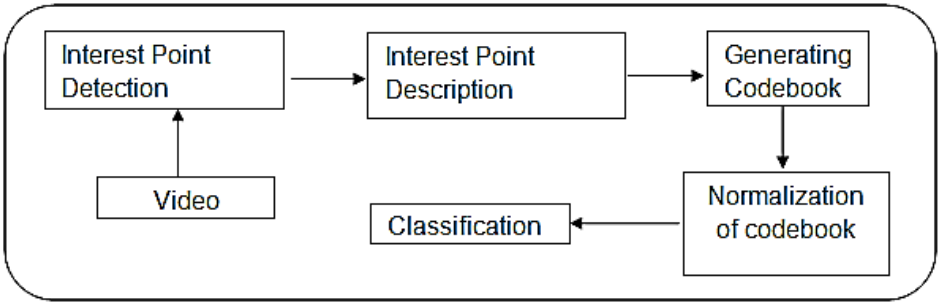

Fig. 1: Steps for Complete Human Action Classification

Interest Point Detection

Interest point detection is basically done by the using the famous SIFT algorithm [18]. Fine tuning of this algorithm was also performed where the threshold value was set to 5. After setting the threshold value to 5 a lot of noisy and unwanted data was removed from the interest points extracted using [12] algorithm. The threshold value targets the changed area of interest points in each frame. This helps us to narrow down the gesture in a video due to change.

The higher the value of the interest point threshold the greater will be the weight of the significance of the interest point which are left behind. Fig 2 displays the interest points after applying the threshold value.

Interest Point Description

The SIFT feature vector has its limitations in detecting the interest points as it cannot differentiate between noise and actual data. For this purpose another enhancement technique has been applied to the interest points using [19] where normalized spatio-temporal Gaussian derivatives are derived from the following equation:

$$L_{x^m,y^n,t^k} = \sigma^{m+n}\tau^k\left(\partial_{x^m,y^n,t^k} g\right) * f$$

The reason for applying Laptev's algorithm is its simplicity and its accurateness in disregarding the points in the frame which is not required for gesture recognition. As seen in figure the first sequence depicts the static background. The second figure displays the successful detection of gait despite the presence of non-stationary background and occlusions.

Generating Codebook

After the detection of all the interest points in our video frame, a vocabulary of visual words is generated. This dictionary is created by sampling the interest points in our training sets. After sampling, clustering is performed on these interest points using k-means clustering. The size of the vocabulary is displayed by the number of clusters and the size of our features [30]. The codebook should be normalized after creation before using bag of words features as the magnitude of the codebook can be dramatically changed after using bag of words [52].

Normalization of Codebook

After a code book is build using k-means clustering and bag of words a normalization technique is used to ensure that the resulting video dataset has the same dimensions. Keeping the various videos dimensions same is the key to having more accurate results. Our proposed algorithm is applied on KTH dataset to whose size ranges from 900 to 1300. Fig displays the accuracy of results on varying video datasets. The figure displays that the best accuracy is achieved on size "xyz"

Three normalization techniques are applied as discussed in [40] and results are shared in the next section:

Normalization Technique N1:

$$p = \frac{p}{\sum_{k=1}^{K} |p_k^2|}$$

Normalization Technique N2:

$$p = \frac{p}{\sqrt{(\sum_{k=1}^{K} |p_k|)}}$$

Proposed Normalization Technique: In the proposed methodology this normalization technique [21] is used:

$$p_{ij} = \frac{p_{ij - \min(p_j)}}{\max(p_j) - \min(p_j)}$$

where p_{ij} is the value of bin number j to be normalized in video number i, max (p_i) and min (p_j) are the maximum and minimum values respectively in bin j over all the videos, now all values are between 0 and 1.

This is the min-max normalization technique. This normalization technique is used to normalize the data from 0 to 1.c. The histograms which are extracted from the visual word vocabulary are treated as two dimensional matrices. The rows are considered as videos and the columns are considered as histograms bins. Then the normalization technique is applied on them.

Bag of words and HMM:

After generation of codebook, bag of words algorithm is performed on the codebook. Object recognition using bag of words has been the most successful, the most studied and the most widely used approaches for object recognition. In this method order less combination of interest point in the local region is represented whereas in general the image is a discrete combination of local regions. After the order less representation of an image the interest points are represented as a histogram over visual vocabulary. This method has shown excellent results in [1][8].

Next comes the HMM part; where feature extraction is performed. Feature extraction is done by BOW and clustering is performed using K-means on visual words. HMM defines the model of temporal behavior between these two states. New interest points in some other image are then mapped into that same state and then prediction algorithm of HMM is performed to determine that on which classification group they belong to.

The HMM models are trained using the histograms which are normalized in the previous step. The testing histograms are also normalized before being fed to the HMMS for classification. Normalization makes the resultant histograms more precise and easy to manage for classification. Hidden Markov Models (HMM) are used for mathematical formulation and solution of this problem. Let $a_{ij}$and $b_j y(t)$denote the probability of a transition from state s_i to state s_j and the emission probability of state s_j generating an observed feature vector $y(t)$. In our problem a_{ij} implies the transition probability from one state to the next, and $b_j y(t)$is the action generated from each state, represented with a vector notation for a set of features. At this point, we make the typical independence assumption for the feature vectors (i.e., features for individual frames are not related). We calculate the probability of a given model producing the observed sequence of feature vectors y_1 to y_T. This sequence must have been generated by a state sequence of length T but because the model is hidden, we do not know the identities of the states. The probabilities for the model generating the observations can be obtained by the joint probability of the observations and any one state sequence, and summing this over all possible state sequences S:

Using the Markov assumption the probability of any particular state sequence is given by the product of the transition probabilities:

$$P(y_1, \ldots\ldots, y_T) = \sum_s P(y_1, \ldots\ldots, y_T | s_1, \ldots\ldots, s_T)\ \mathrm{P}\ (s_1, \ldots\ldots, s_T).$$

RESULTS

In this work KTH dataset is used for training and testing by using the leave out method. KTH dataset is frequently used in the world of action recognition and evaluation. Due to huge level of participation in KTH datasets the data is broken down into and the leave out method has been adopted which is done in [9] where 24 videos are used for training and only one video is used for testing. Then the final recognition rate is determined by averaging out the calculations. The whole dataset contains 598 video clips and each video clip contains only one action. KTH provides a common benchmark to evaluate and compare algorithms. KTH dataset was first time used by Schuldt et al. [57] in 2004 and is one of the mostly widely used datasets until today. It consists of six action sets (hand waving, hand clapping, running, jogging, walking and boxing).

Each action consists of different scenarios such as indoor, outdoor, different clothing and scale changes. Each action is performed by different actors where each run uses 24 videos for clustering and training and one video for testing. Then the final result is computed by taking average.

Tables 1-3 present the confusion matrices of KTH dataset using ‘1-Normalization, ‘2-Normalization, and the proposed normalization technique respectively. The recognition results are presented in the form of average recognition rates. Each entry in the table gives the rate of recognizing of the row action (ground truth) by the column action.

Table 1
Confusion Matrix for First Normalization Method (N1)

	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	0.75	0	0	0.04	0	0
Clapping	0	0.83	0	0.04	0.02	0
Waving	0.15	0.08	0.45	0.09	0.01	0
Jogging	0.25	0.13	0.07	0.26	0.05	0
Running	0	0	0	0	0.72	0
Walking	0	0	0	0.15	0	0.98

Table 2
Confusion Matrix for First Normalization Method (N2)

	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	0.56	0	0	0	0	0
Clapping	0	0.48	0	0.01	0.02	0
Waving	0.05	0.03	0.77	0.05	0	0
Jogging	0.01	0.11	0.07	0.42	0.03	0
Running	0	0	0	0	0.32	0
Walking	0	0	0	0.12	0	0.65

Table 3
Confusion Matrix for Proposed Method of Normalization

	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	1	0	0	0.04	0.02	0
Clapping	0	0.93	0	0.04	0.02	0
Waving	0.15	0.08	0.95	0.09	0.01	0
Jogging	0.25	0.13	0.07	0.98	0.05	0
Running	0	0	0	0	0.94	0
Walking	0	0	0	0.15	0	0.99

Fig. 2: The Effect of Improving SIFT Points Due to Normalization

Fig: 2 presents the effect of fine-tuning the SIFT points due to normalization. The first row displays the SIFT interest points without normalization applied. The second row displays the SIFT interest points with normalization technique applied.

Table 4 shows a comparison between our method and a group of other previously proposed systems that use leave- one-out setup. The results show that for the KTH dataset our result is the best of them.

Table 5 presents a comparison between the overall results (recognition rate) achieved using normalization methods as discussed and also the proposed one. As shown the proposed normalization technique proved positive effort on the performance, and it is worth mentioning that different actors were used for different actions.

Table 4
Comparison with Other Methods

Methods	Accuracy
The proposed method	**98.13**
Bregonzio et al. [49]	94.33
Liu and Shah [48]	94.2
Lin et al. [50]	93.43
Chen and Hauptman [51]	95.83
Niebles et al. [52]	81.5
Tran et al. [53]	95.67
Cao et al. [54]	95.02
Kaaniche and Bremond [55]	94.67
Dollar et al. [10]	81.17
Klaser et al. [56]	91.4
Zhang et al. [47]	91.33
Schuldt et al. [57]	71.72
Fathiand Mori [12]	90.5
Kovashka and Grauman [58]	94.53
Mikolajczyk and Uemura [59]	93.2
Schindler and Gool [35]	92.7
Laptev et al. [22]	91.8
Jhuang et al. [60]	91.7
Gilbert et al. [61]	89.9
Rodriguez et al. [34]	88.7
Nowozin et al. [62]	87
Wong and Cipolla [43]	86.6
Willems et al. [42]	84.3

Table 5
Comparison of Proposed Normalization Method

Normalizations	Accuracy
N1 Normalization	60.30%
N2 Normalization	67.70%
Proposed normalization	95.14%

CONCLUSION

This paper represents human gesture recognition with enhancements to the current algorithm to achieve better results. The algorithm is composed of four stages: detection of interest points using SIFT, feature description using SIFT, creation of visual vocabulary using bag of visual words, and classification using HMM. By employing the normalization enhancement technique the accuracy of the current method has increased by 2% as compared to previous works. Future work can include applying the same algorithm with real time datasets such as in sports, star movements and medicinal field as well. The more complex the datasets the greater the chances are that the normalization would have to be tweaked in order to achieve more accurate results.

REFERENCES

1. Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes. *Inequalities*, 3, 1-8.
2. Al-Ghandi, M., Khan, M.U.G., Zhang, L. and Gotoh, L. (2012). The University of Sheffield and Harbin University at TRECVID 2012: Instance search task at TRECVID, NIST, USA.
3. Agarwal, A. and Triggs, B. (2006). Recovering 3D human pose from monocular images. Pattern Analysis and Machine Intelligence. *IEEE Transactions*, 28.
4. Bay, H., Tuytelaars, T. and Gool, L.V. (2006). SURF: Speeded up robust features. In *Computer vision–European Conference on Computer Vision 2006*, pp. 404-417. Springer Berlin Heidelberg, 2006.
5. Beaudet, P. (1978). Rotationally invariant image operators. In *International Joint Conference on Pattern Recognition*, Vol. 579, p. 583.
6. Blank, M., Gorelick, L., Shechtman, E. Irani, M. and Basri, R. (2005). Actions as space-time shapes. *Tenth IEEE International Conference on Computer Vision*, 2, 1395-1402.
7. Bobick, A.F. and Davis, J.W. (2001). The recognition of human movement using temporal templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23, 257-267.
8. Bobick, A.F. and Wilson, A.D. (1995). Using configuration states for the representation and recognition of gesture. In Proc. *Fifth International Conference on Computer Vision*, 631-636.
9. Gao, Z., Chen, M.Y., Hauptmann, A.G. and Cai, A. (2010). Comparing evaluation protocols on the KTH dataset. In *International conference on human behavior understanding*, 6219, 88-100.

10. Dollar, P., Rabaud, V., Cottrell, G. and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 65-72.
11. Efros, A.A., Berg, A.C., Mori, G. and Malik, J. (2003). Recognizing action at a distance. In *Ninth IEEE International Conference on Computer Vision*, 726-733.
12. Fathi A. and Mori, G. (2008). Action recognition by learning mid-level motion features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1-8.
13. Ferrari, V., Marin-Jimenez, M. and Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2008. 1-8.
14. Freund, Y. and Schapire, R. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14, 771-780.
15. Gorelick, L., Blank, M., Shechtman, E., Irani, M. and Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29, 2247-2253.
16. Haritaoglu, I., Harwood, D. and Davis. L.S. (2000). W4: Real-time surveillance of people and their activities. *IEEE Trans. Pattern. Anal. Mach. Intell*. (TPAMI), 22, 809-830.
17. Harris, C. and Stephens. M.J. (1988). A combined corner and edge detector. In *Alvey Vision Conference*, USA.
18. Vedaldi, A. and Fulkerson, B. And (2008). *VLFeat: An open and portable library of computer vision algorithms*. <http://www.vlfeat.org/>.
19. Laptev. I. (2005). On space-time interest points. *International Journal of Computer Vision*, 64, 107-123.
20. Laptev, I. and Lindeberg, T. (2004). Local descriptors for spatio-temporal recognition. In *Spatial Coherence for Visual Motion Analysis*, 91-103. Springer Berlin Heidelberg.
21. Jayalakshmi, T. and Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. *International Journal Computer Theory Engineering* (IJCTE), 3(1), 89-93.
22. Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *International Conference on Computer Vision and Pattern Recognition*.
23. Lowe. D. (2004). Distinctive image features from scale-invariant key points. *International Journal of Computer Vision*, 60(2), 91-110.
24. Lu, W.L. and Little, J.J. (2006). Simultaneous tracking and action recognition using the pca hog descriptor. The *3rd Canadian Conference on Computer and Robot Vision*.
25. Ke, Y., Sukthanka, R. and Hebert, M. (2005). Efficient visual event detection using volumetric features. IEEE *International Conference on Computer Vision*, 166-73.
26. Moeslund, T.B., Hilton, A. and Kruger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Journal of Computer Vision and Image Understanding*, 104(2-3), 90-126.
27. Parameswaran, V. and Chellappa, R. (2006). View invariance for human action recognition. *International Journal of Computer Vision*, 66, 83-101.

28. Pearson, K. (1901). On lines and planes of closest to systems of points in space. *Philosophical Magazine*, 2, 559-572.
29. Polana, R. and Nelson, R. (1994). Low level recognition of human motion. In *IEEE Workshop on Non-rigid and Articulate Motion*.
30. MacQueen, J.B. (1967). Some methods for Classification and analysis of multivariate observations. Proceedings of *5th Berkeley symposium on mathematical statistics and probability*, 1, 281-97.
31. Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.
32. Ramanan, D., Forsyth, D.A. and Zisserman, A. (2007). Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), 65-81.
33. Ramasso, E., Panagiotakis, C., Rombaut, M., Pellerin, D. and Tziritas, G. (2009). Human shape motion analysis in athletics videos for coarse to one action/activity recognition using transferable belief model. *Electronic Letters on Computer Vision and Image Analysis*, 7, 32-50.
34. Rodriguez, M., Ahmed, J. and Shah, M. (2008). Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *International Conference on Computer vision and Pattern Recognition*.
35. Schindler, K. and Gool, L.V. (2008). Action snippets: How many frames does human action recognition require. In *International Conference on Computer vision and Pattern Recognition*.
36. Scovanner, P., Ali, S. and Shah, M. (2007). A 3-dimensional SIFT descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, pp. 357-360.
37. Senior, A. (2009). An introduction to automatic video surveillance. In *Protecting Privacy in Video Surveillance*. Springer.
38. Sullivan, J. and Carlsson, S. (2002). Recognizing and tracking human action. In *European Conference of Computer Vision*, 629-644. Springer Berlin Heidelberg.
39. Urtasun, R., Fleet, D.J. and Fua, P. (2006). Temporal motion models for monocular and multiview3d human body tracking. *Journal of Computer Vision and Image Understanding*, 104, 157-177.
40. Wang, X., Wang, L. and Qiao, Y. (2012).Comparative study of encoding, pooling and normalization methods for action recognition. *Asian Conference on Computer Vision*, (ACCV), 7726, 572-85.
41. Weinland, D. and Boyer, E. (2008). Action recognition using exemplar-based embedding. In *International Conference on Computer vision and Pattern Recognition*.
42. Willems, G., Tuytelaars, T. and Gool, L.V. (2008). An efficient dense and scale-invariant spatio temporal interest point detector. In *European Conference of Computer Vision*.
43. Wong, S.F. and Cipolla, R. (2007). Extracting spatio-temporal interest points using global information. *IEEE 11th International Conference on Computer Vision*, pp. 1-8.
44. Yamato, J., Ohya, J. and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In *International Conference on Computer vision and Pattern Recognition*.

45. Yilmaz, A. and Shah, M. (2005a). Actions sketch: A novel action representation. *International Conference on Computer vision and Pattern Recognition*.
46. Yilmaz, A. and Shah, M. (2005b). Recognizing human actions in videos acquired by un-calibrated moving cameras. *IEEE 9th International Conference on Computer Vision*, pp. 1-8.
47. Zhang, Z., Hu, Y., Chan, S. and Chia, L.T. (2008). Motion context: A new representation for human action recognition. In *European Conference on Computer Vision*.
48. Liu, J. and Shah, M. (2008). Learning human actions via information maximization. *International Conference on Computer vision and Pattern Recognition*, 1-8.
49. Bregonzio, M., Xiang T. and Gong, S. (2012). Fusing appearance and distribution information of interest points for action recognition. *Pattern Recognition Letters*, 45(3), 1220-34.
50. Lin, Z., Jiang, Z., Davis, L.S. (2009). Recognizing actions by shape-motion prototype trees. *IEEE 12th International Conference on Computer Vision*, 444-451.
51. Chen, M.Y. and Hauptmann, A.G. (2009). MoSIFT: recognizing human actions in surveillance videos. *Technological report*, CMU-CS-09-161, Carnegie Mellon University, 9-161.
52. Niebles, J., Wang, H. and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3), 299-318.
53. Tran, K.N., Kakadiaris, I.A., Shah, S.K. (2011). Modeling motion of body parts for action recognition. *British Machine Vision Conference*, BMVC2011.
54. Cao, L., Liu, Z. and Huang, T.S. (2010). Cross-dataset action detection. *IEEE Conference of Computer Vision and Pattern Recognition*.
55. Kaaniche, M.B. and Bremond, F. (2010). Gesture recognition by learning local motion signatures. *Computer Vision Pattern Recognition*, IEEE 2010, 2745-52.
56. Klaser, A., Marszaek, M. and Schmid, C. (2008). A spatio-temporal descriptor based on 3D-gradients. *British Machine Vision Conference*, BMVC2008, 995-1004.
57. Schuldt, C., Laptev, I. and Caputo, B. (2004). Recognizing human actions: A local SVM approach. *International Conference Pattern Recognition*, ICPR IEEE2004, 3, 32-6.
58. Kovashka, A. and Grauman, K. (2010). Learning a hierarchy of discriminative space-time neighbourhood features for human action recognition. IEEE Computer Vision Pattern Recognition, 2046-53.
59. Mikolajczyk, K. and Uemura, H. (2008). Action recognition with motion-appearance vocabulary forest. *International Conference on Computer vision and Pattern Recognition*, 1-8.
60. Jhuang, H., Serre, T., Wolf, L. and Poggio, T. (2007). A biologically inspired system for action recognition. *International Conference on Computer Vision*, 1-8.
61. Gilbert, A., Illingworth, J. and Bowden, R. (2008). Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *European Conference on Computer Vision* 2008 (pp. 222-233). Springer Berlin Heidelberg.
62. Nowozin, S., Bakır, G.O. and Tsuda, K. (2007). Discriminative subsequence mining for action classification. IEEE *11th International Conference on Computer Vision*, ICCV, 1-8.