ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Volume 2, Special Issue 19, October 2016

Human Action Recognition Techniques- A Survey

Akila.M[#1], Rajeswari.R[#2]

[#] *Department of Computer Applications, Bharathiar University*
Coimbatore, Tamilnadu, India

[1]akila.mca11@gmail.com

[2]rrajeswari@rediffmail.com

***Abstract*— Human action recognition from video is an important research area in the field of computer vision. It is an integral part of surveillance systems, human–computer interactions and various real-world applications. This paper presents a comprehensive review on various methods for human action recognition. It also compares the Speeded-Up Robust Features (SURF) and Histogram of Oriented gradients with Histogram of Optical Flow (HOG/HOF) features based Bag of Words (BOW) approaches for human action recognition. The experimental results were conducted on KTH dataset using a linear Support Vector Machine (SVM) for classification.**

***Keywords*— Human action recognition, HOG/HOF, SURF, Bag of words, SVM.**

I. Introduction

Humans can easily understand actions in a complex scene by using visual system. This field is closely related to other field of studies like motion analysis [1] and action recognition [2].One of the main purposes is to make machines to analyse and recognize human actions using motion information as well as different types of information. There are three important processing stages present in an action recognition system; they are human object segmentation, feature extraction and representation, activity detection and classification algorithms. First, human object is segmented out from the video series. The different features of the human object such as shape, silhouette, colours, poses, and body motions are then properly extracted into a set of features. Thereafter, an action detection or classification algorithm is applied on the features that are extracted in order to recognize various human activities.

The types of human activity are classified under four different categories depending on complexity of actions and number of body parts involved in the action; gestures, actions, interactions, and group activities are the four different types of human activities [3]. Gestures are a collection of movements, made with hands, head or face to show a particular meaning [3]. Actions are a collection of multiple gestures performed by a single person [3]. The walking, waving, jogging, boxing and running are examples of human action categories. Interactions are a collection of human actions of maximum two actors. Group activities are a combination of gestures, actions or interactions where the number of actors is more than two and there may be single or multiple interactive objects [3].

The objective of this paper is to provide a review on various methods for human action recognition, a background of human action recognition and to compare the performance of action recognition systems based on SURF and HOG/HOF features. The rest of the paper is organized as follows: section II reviews previous related work, section III given a background of human action recognition, section IV provides the experimental results and section V gives the conclusion.

II. Existing Action Recognition Methods

This section reviews methods for action recognition in realistic, uncontrolled video data. These methods classified into three categories:

A. Human Body Model Based Methods

Human body model based methods for action recognition use 2D or 3D information on human body parts, such as body part positions and movements.

Johansson[4] has proposed a psychophysical research work based on Human body model methods on visual perception of motion patterns characteristics of living organisms in locomotion. A method has shown that humans can recognize actions from the motion of a few moving light displays attached to the human body, describing the motions of the main human body joints. He has found that between 10 and 12 moving light displays inadequate motion combinations in proximal stimulus evoke an impression of human walking, running and so on.

Yilmaz [5] have proposed an approach for recognition of human actions in videos captured by uncalibrated moving cameras. The proposed approach is based on trajectories of human joint points. In order to handle camera motion and different viewpoints of the same action in different environments, they use the multi-view geometry between two actions and they propose to extend the standard epipolar geometry to the geometry of dynamic scenes where the cameras are moving.

Ali et al.[6] have also proposed an approach based on trajectories of reference joint points. These trajectories are

used as the representation of the non-linear dynamical system that is generating the action, and they use them to reconstruct a phase space of appropriate dimension by employing a delay-embedding scheme. The properties of the phase space are captured in terms of dynamical and metric invariants that include Lyapunov exponent, correlation integral and correlation dimension. Finally, they represent an action by a feature vector which is a combination of these invariants over all the reference trajectories.

B. Holistic Methods

Shape and silhouette information based features were used to represent human body structure and its dynamics for action recognition in videos.

Yamato et al. [7] have proposed an approach using silhouette images and features for action recognition. They extract a human shape mask for each image, calculate a grid over the silhouette, and for each cell of the grid calculate the ratio of foreground to background pixels. Then, each grid representation of an image is assigned to a symbol, which corresponds to a codeword in the codebook created by the Vector Quantization technique. Finally, Hidden Markov Models (HMMs) are applied for action recognition and the model which best matches the observed symbol sequence is chosen as the recognized action category.

Bobick et al. [8] have proposed the idea of temporal templates for action recognition. They extract human shape masks from images and accumulate their differences between consecutive frames. These differences are then used to construct a binary Motion-Energy Image (MEI) and a scalar-valued Motion-History Image (MHI). They indicate the presence of motion, and also represent the pixel intensity of motion at that point. Then, they proposed a recognition method matching temporal templates against stored instances of actions. The MEI and MHI together can be considered as a two component version of a temporal template.

Blank et al. [9] introduced an action model based on space-time shapes from silhouette information. Silhouette information is computed using background subtraction. The authors use the Poisson equation to extract features such as local saliency, action dynamics, shape structure and orientation. Sequences of 10 frames length are then described by a high-dimensional feature vector. During classification, these sequences are matched in a sliding window fashion to space-time shapes in test sequences.

Weinland et al. [10] introduced an orderless 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.

Efros et al. [12] 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 synthetic experiments.

C. Spatio temporal Interest Points

Laptev et al. [13] first introduced the Space–Time Interest Points (STIP) for action recognition by extending the famous Harris detector to video. Their 3D Harris takes into consideration the pixel variations on both space and time. The histogram of Oriented Gradients (HOG) and the Histogram of Optical Flow (HOF) features are then computed in the local neighbourhood of the interest points. The combination of the HOG as a spatial feature representing the local appearances and the HOF as a temporal feature describing the video motions has given promising results.

Dollar et al. [14] observed that sometimes true spatio-temporal corners are rare, even when interesting motion occurs, and might be too rare in certain cases, while enough characteristic motion is still present in other regions. Therefore, they proposed the Gabor detector, which give denser results than the Harris3D. The Gabor detector applies a set of spatial Gaussian kernels and temporal Gabor filters. The final spatio-temporal points are detected as local maxima of the defined response function.

Wong et al.[15] proposed an interest point detector which uses global information, i.e. the organisation of pixels in a whole video sequence, by applying non-negative matrix factorization on the entire video sequence. The proposed detector is based on the extraction of dynamic textures, which are used to synthesize motion and identify important regions in motion. The detector extracts structural information, the location of moving parts in a video, and searches for regions that have a large probability of containing the relevant motion.

Niebles et al. [16] proposed the first extensions to action recognition. For the BoW representation in videos, feature detectors determine a set of salient positions present in the video sequences. Feature descriptors compute a vector representation for the local neighbourhood of a given position. The visual vocabulary (or codebook) is then computed by applying a clustering algorithm (k-means) on feature descriptors obtained from training sequences; each cluster is referred to as a visual word. Descriptors are quantized by assignment to their closest visual word, and video sequences are represented as a histogram of visual word occurrences. Finally a Support Vector Machine (SVM) classifier was used for classification.

Willems et al. [17] proposed the Hessian3D interest point detector, which is a spatio-temporal extension of the Hessian saliency measure for blob detection in images. The Hessian3D detector calculates the Hessian matrix at each interest point and uses the determinant of the Hessian matrix for point localization and scale selection. The detector uses integral video to speed up computations by approximating derivatives with box-filter operations. The detected points are scale-

invariant and dense. A non-maximum suppression algorithm selects joint extrema over space, time and different scales.

Cao et al. [18] proposed an idea for dealing with occlusion and cluttered background by considering the multiple STIP features. In fact, the problems may arise for identifying the actors or distinguishing the actions or tracking of motion field in complex scene where some body parts are occluded by other objects. The associated feature set for action recognition was classified as motion-based feature and appearance-based feature. On top of that, it made use of heterogeneous features such as Hierarchical Filtered Motion Field, sparse feature of histograms of oriented gradient with optic flow (HOG/HOF, histogram descriptors for the space–time volume) and adaptive action detection concept for combining the multiple features by means of Gaussian mixture model (GMM).

Kovashka et al.[19] have proposed to learn a hierarchy of spatio-temporal neighbourhood features in order to capture the most informative spatio-temporal relationship between local descriptors. The main idea is to construct a higher-level vocabulary from new features that consider the hierarchical neighbouring information around each interest point. The method first extracts local motion and appearance features, quantizes them to a visual vocabulary, and then forms candidate neighbourhoods consisting of the words associated with nearby points and their orientation with respect to the central interest point. Descriptors for these variable-sized neighbourhoods are then recursively mapped to higher-level vocabularies, producing a hierarchy of space-time configurations at successively broader scales. And the approach yields state-of-the art performance on the UCF Sports and KTH datasets.

Matikainen et al. [20] have proposed a method for representing spatiotemporal relationships between features in the bag-of-features approach. The authors use both the Spatio-Temporal Interest Points (STIPs) and trajectories to extract local features from a video sequence. Then, they combine the power of discriminative representations with key aspects of Naive Bayes. As the number of all possible pairs and relationships between features is big, they reduce the number of relationships to the size of the codebook. Moreover, they show that the combination of both the appearance and motion base features improves the action recognition accuracy. The main limitation of this technique is that it encodes the appearance and motion relations among features but it does not use information about the spatio-temporal geometric relations between features.

Zhang et al. [22] proposed to model the mutual relationships among visual words by a novel concept named the spatio-temporal phrase (ST phrase). This approach aims to encode rich temporal ordering and spatial geometry information of local visual words. A ST phrase is defined as a combination of k words in a certain spatial and temporal structure including their order and relative positions. A video is represented as a bag of ST phrases which is shown to be more informative than the BoW model.

Yu et al. [21] developed Spatial–Temporal Implicit Shape Model (STISM) for capturing the space–time structure of local sparse features. Due to additive nature of STISM, it can predict multiple actions simultaneously with incomplete observation from segmented video clips. The course of action made use of Multi-class Balanced Random Forest (MBRF) for efficient and discriminative random matching from training set to testing set. Instead of matching all the interest points from training to testing set, the MBRF model brings into focus the interest point pairs; those are falling in same leaf.

Yu et al. [23] provided a real-time solution by considering the spatio-temporal semantic and structural forest for recognizing the actions. It introduced pyramidal spatio-temporal relationship match technique for capturing structural information, those are connected with descriptors. Also, the operating procedure made use of Video FAST for collecting accurate dense interest point in short time sequence. The V-FAST interest point detector provides dense interest point, which has more power to classify the spatio-temporal semantic texton forests (i.e., STF). The spatio-temporal STF generates hierarchical information and codewords by imposing on spatio-temporal patches. It uses kernel k-means forest classifier for efficient classification of vocabularies. Despite that, it used semantic textons for analysing the texture perception and interaction among local space–time elements.

Yan et al. [24] proposed histogram of interest point location (HIPL) algorithm as supplement of bag-of-interest point descriptor for capturing information regarding spatial distribution of STIPs. HIPL is a weaker descriptor, but also has more power to handle large amount of feature vectors. Moreover, the method of working made use of Adaptive Boosting (AdaBoost) and sparse representation (SR) with combination of weighted output classifier for better classification of feature sets. The AdaBoost is a learning algorithm, depicts probability of various classes and makes ensemble of some weak learners.

Chakraborty et al. [25] proposed a novel action recognition algorithm using selective STIPs. The procedure made use of 2D-Harris corner detector with multiple spatial scales in each frame, along with found set of spatio interest points at different scales. The process is made up by removing the unwanted interest points in the background texture by calculating gradient weighting factor. Instead of foreground extraction, the process was involved for suppressing the background by means of non-maxima suppression technique. Thereafter, the system was obtained to selective STIPs with the help of temporal constraint together with matching algorithm. The method of working made use of N-jet descriptor for feature extraction and BOW model for building vocabulary. Finally, SVM was used for action classification and recognition.

Yuan et al. [26] applied the 3D R transform on the interest points based on their 3D locations in order to capture the geometrical distribution of interest points. The 3D R-transform is invariant to geometrical transformation and robust to noise. 2D PCA is then employed to reduce the dimensionality of 2D feature matrix from 3D R transform, obtaining the so-called R features. To encode the features, they combined the R features with the BoW. Finally, they

proposed a context-aware fusion method to efficiently fuse these two features, which is derived from k-nearest neighbour classification approach. The methods described above are summarized in Table I.

TABLE I
RECENT WORKS IN HUMAN ACTION RECOGNITION

Authors	Vocabulary Builder	Classifier	Concept
Laptev et al. [13]	__	Semi-supervised	Implementation of STIPs for compact representation of video data
Dollar et al. [14]	__	__	Recognized behaviour using sparse features in spatio-temporal case
Wong et al. [15]	__	Both	Extracts interest points using global features to identify moving parts
Niebles et al. [16]	BOW	Un-supervised	Unsupervised learning concept for action recognition using spatio-temporal words
Willems et al. [17]	BOW	Supervised	Introduces dense and scale-invariant STIP detector
Cao et al. [18]	State space	Supervised	Action detection using multiple STIP features and focused on cluttered video
Kovashka et al. [19]	Extended BOW	Supervised	Introduces discriminate space–time neighbourhood features for action recognition
Yu et al. [20]	State space	Supervised	Recognized action with temporal semantic and structural forests in real time
Matikainen et al. [21]	BOW	Supervised	Evaluates pairwise spatial and temporal relations for action recognition
Zhang et al. [22]	BOW	Un-supervised	Introduces 4-D local STIP features, combination of dense and intensity information
Yu et al. [23]	BOW	Supervised	Predicts human activities via STIPs detector by introducing forest structures
Yan et al. [24]	BOW	Supervised	Recognized human action using descriptor-based weighted-output classifier for STIPs
Chakraborty et al. [25]	BOW	Supervised	Introduces selective STIPs concept using local descriptor-based approach
Yuan et al. [26]	Mixing BOW	Supervised	Recognized actions using global feature-based STIPs with 3D R transform

III. BACKGROUND

This section describes the action recognition framework that is used in this thesis. The framework consists of four main components: Input Video, Interest Point detection, Bag-of-words representation and SVM classification. The basic steps of bag-of-words model based human action recognition are shown in Fig. 1.

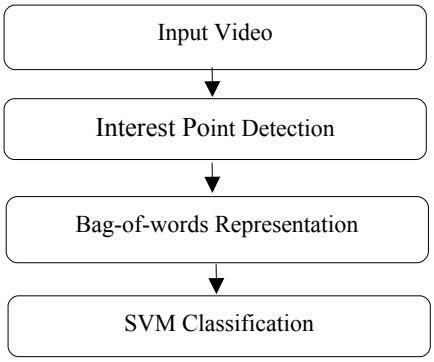

Fig. 1 Action Recognition Framework

A. Input Video

The KTH Dataset[27] contains 6 different actions: walking, jogging, running, boxing, hand waving and hand clapping. The actions are performed by 25 different actors in four different scenarios: outdoors, outdoors with zooming, outdoors with different clothing and indoors. KTH has considerable amounts of intraclass differences. There are differences in duration and somewhat in viewpoint.

B. Interest point detection

The first step is to detect interest points in the video, which are the positions where the features are computed. These points should ideally be located at places in the video where the action is taking place. In this paper Speeded-Up Robust Features (SURF) and HOG/HOF features were implemented.

1) Overview of Speeded-Up Robust Features: The feature vector of SURF creates a grid around the keypoint and divides each grid cell into sub-grids [11] .At each sub-grid cell, the gradient is calculated and is binned by angle into a histogram whose counts are increased by the magnitude of the gradient, all weighted by a Gaussian. These grid histograms of gradients are concatenated into a 64-dimensional vector. The high dimensionality makes it difficult to use this in real time, so SURF can also use a 36-vector of principle components of the 64 vector (PCA analysis is performed on a large set of training images) for a speedup. SURF uses a box filter approximation to the convolution kernel of the Gaussian derivative operator.

2) Overview of Histograms of Oriented Gradients with Histograms of Optical Flow (HOG/HOF): HOG/HOF is the combination of Histograms of Oriented Gradients with

Histograms of Optical Flow [26].A HOG descriptor is computed using a block consisting of a grid of cells where each cell again consists of a grid of pixels. The number *of* pixels in a cell and number of cells in a block can be varied.HOF is computed the same way as HOG but with gradients replaced by optical flow. In [26], HOG is extended to include temporal information, by turning the 2D block in HOG into a 3D volume spanning (x, y, t). This volume is then divided into cuboids that correspond to cells. The sizes of the volume are determined by the detection scale.

C. Bag-Of-Words Representation

The set of local interest point features in a video has to be combined into a representation that enables the comparison with other videos. The most popular method is the bag-of-words representation, where the spatial and temporal locations of the features are ignored. In the first step, the bag-of-features model builds a visual vocabulary, called codebook. The codebook is generated using local features extracted from the training videos. Local features extracted from the testing videos are not used in the process of creating the codebook. Typically, the codebook is generated using the k-means algorithm. After generating the visual vocabulary (codebook), every video can be represented by the bag-of-features model. The bag-of-features model represents a video sequence by assigning its features to the nearest elements of the created visual vocabulary, i.e. to the nearest cluster centers. Finally, normalize the histogram representation so that the video size does not significantly change the bag-of-features magnitude.

D. SVM Classification

Support Vector Machines (SVMs) are among the most prominent machine learning algorithms that analyze data and recognize patterns [27]. They are widely used for classification and regression. SVMs are one of the most robust and accurate Machine Learning methods. The aim of the SVMs is to find the optimal hyperplane which separates two classes of data. Depending on the nature of the data, such a separation might be linear or non-linear. SVMs belong to the supervised learning algorithms. It means that they use training samples, where each training sample is a pair of an input object (typically a vector) and a desired output value (class label). The SVMs analyze the training data and build an inferred function that can be used to correctly determine the class label for an unseen input object.

IV. Experimental Results

In this work, a comparison of human recognition methods using two interest point detectors is carried out. The interest point detectors used are SURF and HOG/HOF features. The default settings are used for parameters like number of temporal and spatial scales and detector sensitivity. Bag of Words model based on these features are constructed which are then classified using SVM. The detected interest points are shown in Fig. 2.

For the datasets, a number of 4000 vocabularies are built to find the best strategy for vocabulary generation. Each individual experiment utilizes the vocabulary that provides the best result for that specific experiment. The measure used for comparison is ‘‘mean average precision”. Average precision is the average of all true positive percentages across classes. The classifier used is a Support Vector Machine (SVM) and the implementation used is libsvm [28]. Fig. 3 displays the mean average precision for SURF and HOG/HOF features.

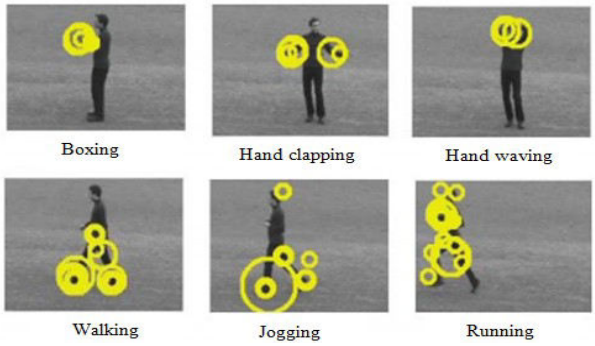

Fig. 2 Detected Interest Points on KTH Dataset

The Mean average precision measures for different actions based on detected features are shown in Table II.

TABLE III
Mean Average Precision Measures

Features / Actions	SURF		HOG/HOF	
	Training Mean AP	**Testing Mean AP**	**Training Mean AP**	**Testing Mean AP**
Walking	0.992	0.714	0.934	0.505
Jogging	1.000	0.375	0.847	0.282
Running	0.964	0.611	0.856	0.289
Boxing	0.676	0.267	0.964	0.671
Hand waving	1.000	0.494	0.964	0.223
Hand clapping	1.000	0.658	1.000	0.286
Mean AP for KTH Dataset	0.939	0.520	0.927	0.376

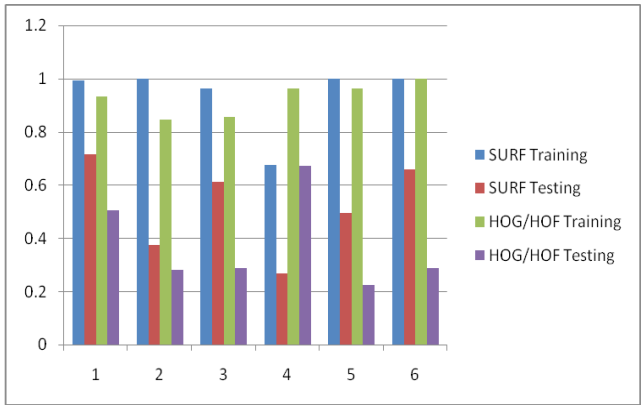

Fig. 3 Mean Average Precision for SURF and HOG/HOF features

Based on the results shown, SURF based features produce Mean Average Precision as 0.520 which comparatively gives good result than HOG/HOF features.

V. Conclusion

This paper presented a comprehensive study on human action recognition methods. There has been an increase in the number of Spatio Temporal Interest Point based approaches for human action recognition. Action recognition becomes very difficult for multiple moving objects in the presence of shadow, illumination changes in the scene. In this paper Speeded-Up Robust Features (SURF) and Histogram of Oriented gradients with Histogram of Optical Flow (HOG/HOF) features based bag-of-words model for human action recognition are also compared using KTH dataset.

References

[1] Aggarwal.J.K and Cai.Q,"Human Motion Analysis :A Review", IEEE Proceedings of Nonrigid and Articulated Motion Workshop vol. 73, no.3, pp.428–440, 1999.

[2] Wu J, Hu D , Chen F, "Action recognition by hidden temporal models",Vis. Comput. 30(12), pp.1395–1404, 2003.

[3] Aggarwal J.K., Ryoo M.S, "Human activity analysis: a review", ACM Compuer. Surv. (CSUR). 43(3), 16:1–16:43,2011.

[4] Johansson G,''Visual perception of biological motion and a model for its analysis'', Perception & Psychophysics, vol.14,pp.201–211, 1973

[5] Yilmaz A and Shah M, ''Recognizing human actions in videos acquired by uncalibrated moving cameras'',Computer Vision, Tenth IEEE International Conference, vol. 1, pp 150–157,2005

[6] Ali S, Basharat A and Shah M, ''Chaotic invariants for human action recognition'', Computer Vision ICCV,IEEE 11th International Conference, pp 1–8, 2007.

[7] Yamato J, Ohya J and Ishii K, ''Recognizing human action in timesequential images using hidden Markov model",Computer Vision and Pattern Recognition,Proceedings CVPR '92., IEEE Computer Society Conference, pp. 379–385, 1992.

[8] Bobick A.F and Davis J.W, ''The recognition of human movement using temporal templates", IEEE Pattern Analysis and Machine Intelligence,vol. 23, pp.257–267, 2001.

[9] Blank M, Gorelick L, Shechtman E, Irani M and Basri R, "Actions as space-time shapes", Tenth IEEE International Conference, vol. 2, pp 1395–1402. 2005.

[10] Weinland D and Boyer E, "Action recognition using exemplar-based embedding," Computer Vision and Pattern Recognition, IEEE Conference, pp. 1–7, 2008.

[11] Pinto B, and Anurenjan P.R, ''Video stabilization using Speeded Up Robust Features", IEEE International Conference on Communications and Signal Processing , pp. 527 531, 2011.

[12] Efros A, Berg A, Mori G, and Malik J, "Recognizing action at a distance," Computer Vision,Ninth IEEE International Conference, pp. 726–733, 2003.

[13] Laptev I, "On space-time interest points", International Journal on Computer Vision, vol.64, pp.107–123, 2005.

[14] Dollar P, Rabaud V, Cottrell G and Belongie S, ''Behavior recognition via sparse spatio-temporal features", 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72,2005.

[15] Wong S and Cipolla R, ''Extracting spatiotemporal interest points using global information", 11th IEEE International Conference of Comp Vis, pp. 1–8,2007.

[16] Niebles J, Wang H and Fei-Fei L, ''Unsupervised learning of human action categories using spatial–temporal words", International Journal of Computer Vision. vol.79, pp.299–318, 2007.

[17] Willems G,Tuytelaars T and Gool L.V,''An efficient dense and scale-invariant spatio-temporal interest point detector", Computer Vision ECCV,vol.5303,pp.650–663,2008.

[18] Cao L,Tian Y.L,Liu Z,Yao B and Huang T.S, ''Action detection using multiple spatial–temporal interest point features", International Conference on Multimedia and Expo, pp. 340–345,2010.

[19] Kovashka A, Grauman.K, ''Learning a hierarchy of discriminative space-time neighborhood features for human action recognition",IEEE Conference on Computer Vision and Pattern Recognition, pp. 2046–2053,2010.

[20] Yu T.H, Kim T.K and Cipolla R, ''Real-time action recognition by spatiotemporal semantic and structural forests", Proceedings of the British Machine Vision Conference, pp. 52.1–52.12,2010.

[21] Matikainen P, Hebert M and Sukthankar R, ''Representing pairwise spatial and temporal relations for action recognition", Proceedings on 11th European conference of the Computer vision, pp. 508–521,2010.

[22] Zhang H and Parker L.E, "4-Dimensional local spatio-temporal features for human activity recognition", Proceedings of IEEE International Conference on Intelligent Robots and Systems, pp. 2044–2049,2011.

[23] Yu G, Yuan J, Liu Z, "Predicting human activities using spatiotemporal structure of interest points", Proceedings of the 20th ACM international conference on Multimedia, pp. 1049–1052, 2012.

[24] X.Yan and Y.Luo, "Recognizing human actions using a new descriptor based on spatialtemporal interest points and weighted-output classifier", Journal of Neurocomputing, vol.87, pp.51–61, 2012.

[25] Chakraborty B, Holte M.B, Moeslund T.B and Gonzalez J,"Selective spatio-temporal interest points", Special issue on Semantic Understanding of Human Behaviors in Image Sequences, vol. 116, pp. 396–410,2012.

[26] Yuan C, Li X, Hu W, Ling, H,Maybank S, "3D R transform on spatio-temporal interest points for action recognition" In CVPR,2013

[27] Schuldt C, Laptev I and Caputo B, "Recognizing human actions: a local SVM approach", 17th International Conference on ICPR 3, pp. 32–36,2004.

[28] Chang C and Lin C, "LIBSVM : A Library for Support Vector Machines," pp. 1–39, 2013.