SPATIOTEMPORAL SALIENCY AND SUB ACTION SEGMENTATION FOR HUMAN ACTION RECOGNITION

Abhishek Babu
Dept. of Computer Science and Engineering
TKM College of Engineering
Kollam
abhisceedee@gmail.com

Shyna A
Dept. of Computer Science and Engineering
TKM College of Engineering
Kollam
shyna@tkmce.ac.in

Abstract—**Human Action Recognition is a significant and challenging field of interest in Research and Industry. In this paper, the Selective Spatiotemporal Interest Points (Selective STIPs) are extracted from the input video and is labeled using a dictionary. The actions are segmented into sub-actions, and then the temporal and spatial structure is captured. The segmentation is done on the basis of interest point density. The spatial and temporal relationships between the labeled STIPs is represented using Space Salient and Time Salient directed graphs respectively. Time Salient pairwise feature (TSP) and Space Salient pairwise feature (SSP) is computed from corresponding directed graphs. The Selective STIP suppresses the background STIPs and detects more robust STIPs from the actors which improves performance of recognition. The Bag-of-Visual Words model combined with TSP and SSP for human action classification provides a more promising result.**

Index Terms— **Selective STIP, Bag-of-Visual Words, Time Salient Directed Graph, Space Salient Directed Graph, SVM**

I. Introduction

In last two decades, we have seen a lot of advancements in the study of human action recognition (HAR). HAR plays a key role in human-machine interaction, the Microsoft Kinect is such a device which uses RGB-D camera which can also capture depth information. This is achieved by incorporating a depth sensor into the device.

Depth sensors can increase the accuracy of HAR device, but it has its own limitations. Depth sensors increase the computation overhead and has range limitations making it unsuitable for low computing devices such as an intelligent surveillance camera. The intelligent surveillance camera is used to monitor the human activity and alert when an abnormal activity is identified. HAR can also be applied for Content based video retrieval.

Recognizing human actions with high precision is a challenging task due to problems such as background interference, body occlusion, scale changes, change of view point, environment changes, moving camera, zooming, different action performance speed, and different anthropometry of the actors and their movement style variations. The actions with similar sub-actions also increases the difficulty of action classification.

The human activity recognition is classified into unimodal and multimodal methods according to the nature of sensor data deployed. Unimodal methods uses a single source, such as an image for recognizing human activities, and they are further categorized as: (i) space-time, (ii) stochastic, (iii) rule-based, and (iv) shape based methods. Multimodal methods combine features collected from different sources and are classified into three categories: (i) affective, (ii) behavioral, and (iii) social networking methods [1].

The recent interest of research is in Space-Time approach, which uses the spatial-temporal features for action classification. The challenge in this approach is to find robust Spatiotemporal Interest points (STIPs) by suppressing unwanted background STIPs which may lead to wrong classification of action. The interest points can be a corner points or end points of a line or an isolated point or point of curve. There are various methods available for the detection of STIPs. On the basis of feature sets, the STIP detector is grouped into dense and sparse feature-based detector [2]. A dense feature detector densely covers the video content, e.g. Hessian detector, Dense sampling, V-FAST. On the other hand, a sparse (local) feature is a subset of the total feature vector (i.e., large and unbound), e.g. Harris 3D, Cuboid detector. The spatiotemporal-based descriptors are classified as local descriptor and global descriptor. A local descriptor captures only local or static information (i.e., color, posture, texture etc.), e.g. ESURF, Cuboid descriptor, N-jet, while a global descriptor captures the global or dynamic information (i.e., scale changes, illumination changes, speed variation, phase variation, etc.) of a video, e.g. HOG 3D, HOG/HOF. The shape or edge relevant data of objects are used for determining the local features of a video. However, the global information points towards the description of flow or motion of a video. By selecting the most efficient detection and description algorithm, we can improve the performance of action recognition.

The Bag-of- Visual Words model does not consider the spatial and temporal structure, therefore a relationship between the STIPs can be used to incorporate the spatiotemporal structure to the existing BoVW model [3]. By sub-action segmentation, the same sub-class is classified into a sub-section for a particular action [4].

II. LITERATURE SURVEY

Bag-of-visual words (BoVW) describes an action by order less bag of features [5,6] .The method does not work in distinguishing actions with similar sub actions, where the order of the sub actions plays a significant role. In Dynamic Word Model, the word distribution changes over time [7]. Bag-of-words model ignores spatiotemporal structure information. The action is represented by the histogram of feature words BoW model. Guha and Ward [8] explores the sparse representations for recognizing human activity. An over complete dictionary is constructed using the dictionary elements, i.e. spatiotemporal descriptors .Laptev [9] applied Histogram of Oriented Gradient (HOG) as STIP descriptor. Laptev also combined HOG and Histogram of Optical Flow (HOF) to form a new descriptor which is more efficient than the individual methods. Laptev and Lindeberg extended Harris Corner detector to 3D [9]. The action is represented as a cloud of STIPs [10]. Dollar applied gabor filter on spatial and temporal dimension [11]. Klaser extended HOG to 3D [12]. Willems extended SURF feature to 3D [13]. Scovanner extended SIFT feature to 3D [14].Wang used dense Optical Flow trajectories to describe the motion patterns of an object [15]. Wu used visual features and Gaussian Mixture Model to represent spatiotemporal context distribution [1]. Vrigkas used Gaussian mixture model to cluster the motion trajectories, the motion trajectories represents human action, and the action labeling is performed using a nearest neighbor classification scheme [1].

Niebels used probabilistic Latent Semantic Analysis (pLSA) model to learn probability distribution of words. Wong extended pLSA to pLSA with Implicit Shape Model (pLSA-ISM) [16] .Yuan proposed a 3D discrete Randon Transform to capture distribution of 3D points [17]. Zhao developed Local Binary Pattern on Three Orthogonal Planes [18,19]. Shao developed Extended LBP-TOP and Extended CSLBP-TOP [20].

Ryoo and Aggarwal introduced spatio-temporal relationship match (STR match) which considers spatial and temporal relationships to recognize human action [18,16]. Seo and Milanfar used space-time locally adaptive regression kernel and the matrix cosine measure for representation of human actions [21].

Yu (2012) proposed a propagative point-matching approach which uses random projection trees and has the ability to label the data in unsupervised manner [22]. Jain et al. proposed a new motion descriptor called Divergence-Curl-Shear descriptor and used motion compensation techniques to recognize atomic actions [23]. Gaidon et al. used short tracklets and proposed unsupervised learning method recognizing human activity from it [24]. The hierarchical clustering algorithm represents the videos with an unordered tree structure and compares all tree-clusters to identify the activity. Li et al. (2011) explored temporal information in a video sequences to identify the human action [25]. Messing et al. (2009) extracted motion features from the video sequence [26]. These extracted features were tracked by their velocities, and velocity history of the trajectories is learned by a generative mixture model which classifies each video clip. Tran et al. (2014a) proposed a scale and shape invariant method for localizing spatiotemporal events and uses a sliding window technique to track spatiotemporal path in video sequences [27].

Support Vector Machines uses local descriptors of fixed length as input for the classifier [28]. Relevance Vector Machines (RVM) is a probabilistic variant of SVM. Sparser set of support vectors is obtained by training a RVM. Oikonomopoulos used RVM for action recognition in his method [29].

III. PROPOSED SYSTEM

The main goal of the research is to develop an accurate action recognition system. The proposed system is based on spatiotemporal interest points. Compared to work by Liu [3], the proposed system automatically segments the actions into sub-actions and does not use any fixed threshold values. The proposed work uses a variable threshold that can segment an action into sub-actions such that sub-actions in action segment of similar actions have same nature and thus can be used for the purpose of comparison. At first, extract the spatio-temporal interest points from the input video clip. We use Selective STIP detector developed by Chakraborty for extracting the interest points [30]. The Selective STIP detector outperforms the other state-of-the-art methods and has better efficiency than the detection method used by Liu[3] in his work. Table 1 shows STIP detection ratios and clearly documents the superior performance of Selective STIP over other methods.

The HOG feature is used for describing the STIPs. The STIPs extracted are labeled based on a predefined dictionary. The actions in the input video is segmented into sub-actions based on the intensity of the interest points. The sub-actions are small period of actions. Each action class will contain same number of sub-actions. STIP density is used for action segmentation. The number of frames in a segment is less where the STIP density is high. Further, the motion range is calculated to eliminate speed and range differences between instances.

TABLE I. STIPs DETECTION RATIOS (%)

Method	**Dataset**	
	MSR I	***Multi –KTH***
Chakraborty	76.21	90.34
Laptev and Lindeberg	18.73	48.16
Dollar	21.36	16.03
Willems	24.02	20.24

The Time Salient directed graph for each sub actions is generated. The Time Salient Pairwise feature (TSP) and Space Salient Pairwise feature (SSP) is then computed for each sub actions. The combination of TSP, SSP and BoVW is used as the feature to train the SVM. The training is done by using the action videos available in KTH dataset. Each steps is explained in detail in below Sections.

A. Selective Spatio-temporal Interest Points

Selective STIP detector detects the Spatial Interest Points (SIPs) from the video and suppresses the unwanted background SIPs. More robust STIPs for actor is obtained by imposing local and temporal constraints. Selective STIP detects dense STIPs at the motion region without affected by the complex background. The HOG descriptor of Selective STIPs is used as the feature vector. From the training videos, we collect all the feature vectors and applies k-means clustering algorithm to generate a dictionary. The dictionary uses the K cluster centers to label all the Selective STIPs obtained from the input video. The descriptor of Selective STIPs is labeled to the nearest cluster center in the dictionary.

B. Action Segmentation

The action segmentation is efficient if the sub-actions in the same sub-section of same action class are of the same type, ignoring the speed difference between different actors. This is achieved by two stage segmentation.

In the first segmentation stage, the entire video is chopped into clips with approximately equal number of interest points in all clips. The segmentation by interest point density is better than by the one using the number of frames, as it ensures motion information in all clips.

$$D_c = \sum_{i=1}^{N}(n_i)/N_c \qquad (1)$$

The number of interest points in a clip (D_c) is determined by Equation 1. N is the total number of frames in the input video. N_c is the clip number and n_i represents the number of interest points in i^{th} frame.

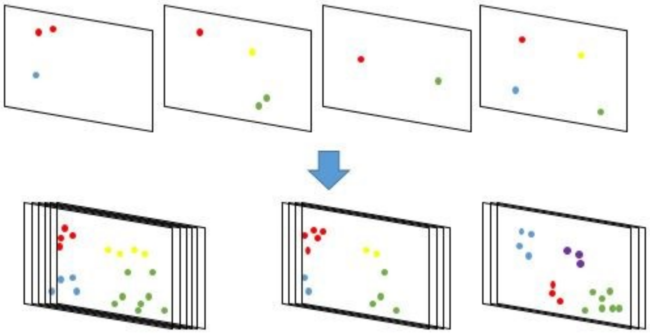

Fig. 1. Combining the frames to form N_s sub-actions.

Before second stage of segmentation, the histogram of visual words in each clips needs to be computed. Histogram of k^{th} clip is represented by h_k . In second stage of segmentation the motion range is measured with χ^2 distance between neighboring clips as shown in Equation 2. The value of k ranges from 2 to N_c. X stores the distance between the adjacent clips. Fig. 1 shows the segmentation of actions into sub-actions.

$$X(k-1) \ = \chi^2(\ h_{k-1},\ h_k) \qquad (2)$$

The threshold T is used to segment clip series. The value of T is determined by Equation 3. Ns is the number of sub actions in an action. The value of Ns is constant to all actions and is selected from the action class having the least sub-actions.

$$\mathrm{T} = \sum X \ / \ Ns \qquad (3)$$

The starting clip and ending clip is selected in such a way that the summation of X in each action segment is greater than the threshold T. After segmentation, the time salient directed graph is computed for each segments.

C. Time Salient Directed Graph

The set of all STIPS of an action sequence in represented by **S**. The points in **S** is represented by the set {x,y,t,label},where x and y represents the geometric positions of point, t represents the frame number and label stores the label assigned to each point from the dictionary. For each sub-action segment a Time Salient Directed Graph (TSD) is computed. The STIPs in a segment forms the vertices of TSD graph. Two points (a,b) in a segment is said to be an edge of TSD if a and b have different labels and the difference of x or y coordinates of a and b is greater than some threshold T_x and T_y respectively. If any of the one condition fails then the pair of points is discarded. The TSD gives a relationship between the different labels in a single sub-action segment.

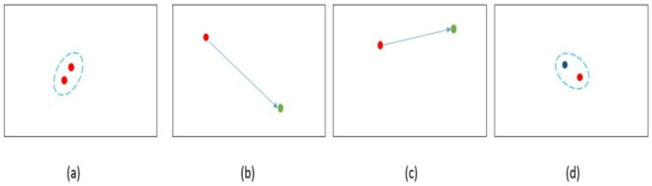

Fig. 2. Assigning directions to label points. (a) and (d) are ignored from assigning directions whereas in (b) and (c), the directions gets assigned.

The direction of the edges is the next concern and vertical relationship is given more priority than horizontal.

```
Algorithm 1: Assigning edge direction
    1. $\Delta x = abs(x_a - x_b)$
    2. $\Delta y = abs(y_a - y_b)$
    3. if $\Delta x \geq T_x$
        a. if $x_a < x_b$ then a → b else b → a
    4. elseif $\Delta y \geq T_y$
        a. If $y_a < y_b$ then a → b else b → a
```

Fig. 2 shows the assignment of directions to label points. Since there is N_s sub-action for an action, we get N_s number of TSDs. A distribution map **M** is created from these TSDs. The distribution map is a **KxK** relation between the labels. The i^{th} row in distribution map **M** represents the sum of out-degree from label i points to other labeled points in all TSDs. The i^{th} column represents the sum of in-degree to label i points from the other labeled points in all TSDs. The count of the frames in which the label i point occurs is represented by C(i).

$$TSP_{in}(i) = \frac{\sum_{j=1}^{K} M(j,i)}{\sum_{j=1}^{K} C(i).C(j)} \qquad (4)$$

$$TSP_{out}(i) = \frac{\sum_{j=1}^{K} M(i,j)}{\sum_{j=1}^{K} C(i).C(j)} \qquad (5)$$

$TSP_{in}(i)$ represents probability of label i being the ending point and $TSP_{out}(i)$ represents probability of label i being the starting point in the TSDs. The Time Salient Pairwise feature TSP is represented as below.

$$TSP = \{[TSP_{in}(i)]^{K}_{i=1}, [TSP_{out}(i)]^{K}_{i=1}]\} \qquad (6)$$

D. Space Salient Directed Graph

The Space Salient Directed Graph SSD gives a relationship between the same labels across the sub-actions. In a single sub action segment, there may be several points with same labels and occurs nearby to each other. Here we take the mean of same labeled points and considers it as the unique point for a label in a sub-action segment. The SSD is a set of vectors connecting the same labeled points across the sub-action segments. If $pt_i = (x_i,y_i,s_a)$ and $pt_j = (x_j,y_j,s_b)$ are two points of same label in two different sub-action segment s_a and s_b, then $SSP(pt_i,pt_j) = (x_i-x_j,y_i-y_j,s_a-s_b).\delta(s_a,s_b)$,where $\delta(s_a,s_b)$ is +1 if $s_a < s_b$ and it is -1 when $s_b < s_a$. Thus a vector bank which connects all the same labeled points across the action segments is formed.

The vectors collected from the training videos are initially clustered into K_2 cluster points. The vectors obtained from the testing video is then labeled to the nearest from the K_2 labels. The Space Salient Pairwise feature SSP is the histogram of the labeled vectors present in the video.

E. Human Action Classification

The block diagram for training and testing of human activity is shown in Fig. 3 and Fig. 4, respectively.

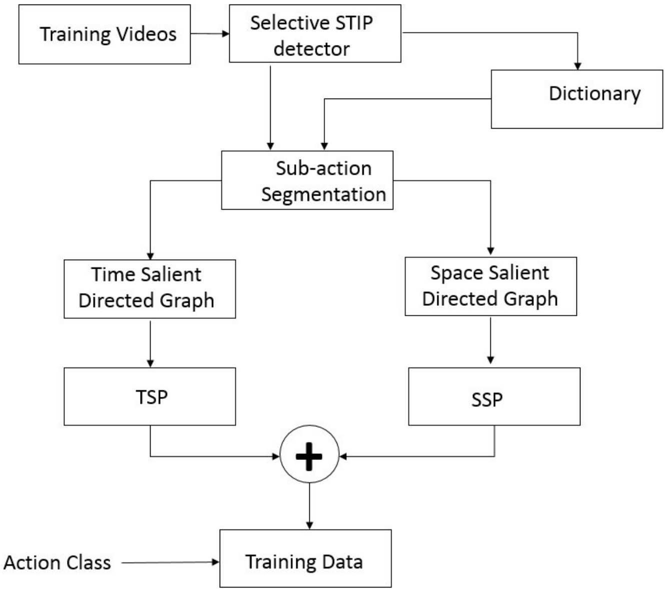

Fig. 3. Block diagram for training of Classifier.

The training data is labeled with the action classes explicitly. In Fig. 3 and Fig. 4 the features extracted are TSP and SSP. The efficiency of the proposed system can be tested in four cases, i.e. by using BoVW, TSP+ BoVW, SSP+ BoVW or by TSP + SSP + BoVW as feature set.

In Bag of Visual Words model we find the histogram of labels from all the STIPs. The clustering is done through k means and the best value of k for each method can be obtained by testing. The TSP+SSP+BoVW feature from training video is used for training the non-linear SVM classifier with a homogeneous kernel.

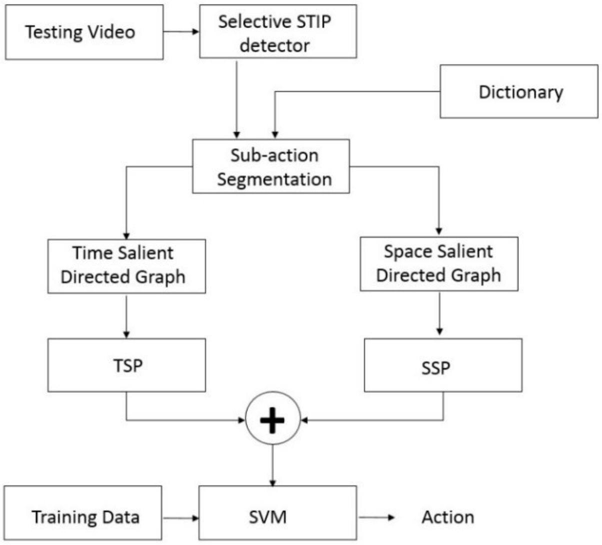

Fig. 4. Block diagram for human action recognition.

IV. EXPERIMENTAL RESULT

The dataset we are using for action recognition in our experiment is KTH dataset, which contains a total of 600 videos of 25 persons performing 6 actions repeated to 4 times with different backgrounds. The six actions performed are “walking”, “jogging”, “boxing”, “running”, “hand waving” and “hand clapping”. The system is implemented in MATLAB R2013b and the efficiencies are calculated on laptop with Intel® Core™ i5-3210M CPU @ 2.50GHz and 4GB of RAM. The experiment is conducted with value of $T_x = 5$ and $T_y = 5$. The number of clusters K = 200 and K_2 = 100.The BoVW model we use k means clustering with value of k = 900.

TABLE II. COMPARISON WITH RELATED WORK ON KTH

Method	Accuracy (%)	
	M Liu et al.	*Proposed Method*
BoVW	93.83	93.83
TSP + BoVW	94.50	96.10
SSP + BoVW	95.67	95.85
SSP + TSP + BoVW	95.83	96.47

Table II compares the efficiency of proposed system with the method proposed by M. Liu et al. [3]. The exeperiment result shows that the proposed system outperforms the existing method for human action recognition. The STIP detection ratio is given in Table I and it also show that the Selective STIPs out performs other STIP detection methods.

V. CONCLUSION

Comparing with the method proposed by M. Liu et al. [3], our novelty lies in the use temporal information of a sub-action and spatial relation between the sub-actions. The sub-action segmentation classifies the same sub-class into same sub-section for a particular action. Therefore in each sub-class we get similar sub-action. In the proposed method we consider the spatiotemporal relationship between Selective STIPs across the sub-action segments. TSP feature represents the relationship between different labeled points in a sub- action segment. SSP feature represents the spatial relationship between the same labeled points across the sub-actions. The combination TSP+SSP+BoVW feature is used as input to non-linear SVM classifier with homogeneous kernel. The proposed system outperforms the existing method as it uses efficient Selective STIPs and considers the relationship between the sub-actions in an action segment along with the BoVW model.

REFERENCES

[1] Vrigkas M, Nikou C and Kakadiaris IA (2015) “A Review of Human Activity Recognition Methods”. Front. Robot. AI 2:28.

[2] Dawn, D.D. and Shaikh, S.H., 2016. “A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector” The Visual Computer, 32(3), pp.289-306

[3] Liu, Mengyuan, et al. "Salient pairwise spatio-temporal interest points for real-time activity recognition." Caai Transactions on Intelligence Technology 1.1 (2016): 14-29.

[4] Liu, Hong, et al. "Sequential Bag-of-Words model for human action classification." CAAI Transactions on Intelligence Technology 1.2 (2016): 125-136.

[5] Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: A local SVM approach." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 3. IEEE, 2004.

[6] Dollár, Piotr, et al. "Behavior recognition via sparse spatio-temporal features." Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. IEEE, 2005.

[7] Prest, Alessandro, Cordelia Schmid, and Vittorio Ferrari. "Weakly supervised learning of interactions between humans and objects." IEEE Transactions on Pattern Analysis and Machine Intelligence 34.3 (2012): 601-614.

[8] Guha, Tanaya, and Rabab K. Ward. "Learning sparse representations for human action recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence 34.8 (2012): 1576-1588.

[9] Laptev, Ivan. "On space-time interest points." International journal of computer vision 64.2-3 (2005): 107-123.

[10] Bregonzio, Matteo, Shaogang Gong, and Tao Xiang. "Recognising action as clouds of space-time interest points." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[11] Dollár, Piotr, et al. "Behavior recognition via sparse spatio-temporal features." Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. IEEE, 2005.

[12] Klaser, Alexander, Marcin Marszałek, and Cordelia Schmid. "A spatio-temporal descriptor based on 3d-gradients." BMVC

2008-19th British Machine Vision Conference. British Machine Vision Association, 2008.

[13] Willems, Geert, Tinne Tuytelaars, and Luc Van Gool. "An efficient dense and scale-invariant spatio-temporal interest point detector." Computer Vision–ECCV 2008 (2008): 650-663.

[14] Scovanner, Paul, Saad Ali, and Mubarak Shah. "A 3-dimensional sift descriptor and its application to action recognition." Proceedings of the 15th ACM international conference on Multimedia. ACM, 2007.

[15] Wang, Heng, et al. "Action recognition by dense trajectories." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[16] Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. ACM Computing Surveys (CSUR), 43(3), 16.

[17] Yuan, Chunfeng, et al. "3D R transform on spatio-temporal interest points for action recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[18] Ryoo, Michael S., and Jake K. Aggarwal. "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities." Computer vision, 2009 ieee 12th international conference on. IEEE, 2009.

[19] Poppe, Ronald. "A survey on vision-based human action recognition." Image and vision computing 28.6 (2010): 976-990.

[20] Shao, Ling, and Riccardo Mattivi. "Feature detector and descriptor evaluation in human action recognition." Proceedings of the ACM International Conference on Image and Video Retrieval. ACM, 2010.

[21] Seo, Hae Jong, and Peyman Milanfar. "Training-free, generic object detection using locally adaptive regression kernels." IEEE Transactions on Pattern Analysis and Machine Intelligence 32.9 (2010): 1688-1704.

[22] Yu, Gang, Junsong Yuan, and Zicheng Liu. "Propagative hough voting for human activity recognition." Computer Vision–ECCV 2012 (2012): 693-706.

[23] Jain, Mihir, Herve Jegou, and Patrick Bouthemy. "Better exploiting motion for better action recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[24] Gaidon, Adrien, Zaid Harchaoui, and Cordelia Schmid. "Recognizing activities with cluster-trees of tracklets." BMVC 2012-British Machine Vision Conference. BMVA Press, 2012.

[25] Li, Binlong, et al. "Activity recognition using dynamic subspace angles." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011

[26] Messing, Ross, Chris Pal, and Henry Kautz. "Activity recognition using the velocity histories of tracked keypoints." Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009.

[27] Tran, Du, Junsong Yuan, and David Forsyth. "Video event detection: From subvolume localization to spatiotemporal path search." IEEE transactions on pattern analysis and machine intelligence 36.2 (2014): 404-416.

[28] Jhuang, Hueihan, et al. "A biologically inspired system for action recognition." Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. Ieee, 2007.

[29] Oikonomopoulos, Antonios, Ioannis Patras, and Maja Pantic. "Spatiotemporal salient points for visual recognition of human actions." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36.3 (2005): 710-719.

[30] Chakraborty, Bhaskar, et al. "Selective spatio-temporal interest points." Computer Vision and Image Understanding 116.3 (2012): 396-410.