New Approach For Moving Point Detection: Application To Video Space-Time Description

Olfa Ben Ahmed(1,2) /Mahmoud Mejdoub(1)
1: REsearch Group on Intelligent Machines
Sfax, Tunisia
olfa.ben-ahmed@labri.fr, mah.mejdoub@gmail.com

Chokri Ben Amar (1)
2 : LABRI UMR 5800 CNRS/Université Bordeaux 1
France, Bordeaux
Chokri.benamar@ieee.org,

Abstract— **Local space time points detection and description have recently emerged as a major research in based video analysis content in video surveillance and event detection applications. Several works extend tow dimensional descriptor to the temporal dimension. Most of the existing methods consider the video as a spatio-temporal volume and then describe the volumetric region around the salient point in 3D. However, this representation requires a high operational complexity. In this works we propose a new approach to describe motion using a simple 2D representation of the video. Our method is based on tracked feature points in image sequence. The main challenge in motion description is how to detect the local motion information. In this paper we aim to describe motion around moving point without the need to extend them on the 3D dimension. To show the efficiency and accuracy of our approach, we perform action recognition experiments on the KTH and Weizmann databases using the bag of words approach. We have obtained impressive results for action recognition.**

Keywords-component; Moving point, video, Accordion, Bag of words, space-time description

I. INTRODUCTION

Extracting features from a video sequence is the first step of many video processing applications, including video surveillance, indexing of video archives and event detection. Actually, human action detection and recognition is an active research area. Since, most interest activities in videos are characterized by motion variations of image structures over time, several methods and techniques for robust motion detection and appearance description in video have been proposed in recent years. The authors on [1] constructed Motion-Energy Images (MEI) and motion history images (MHI) as temporal templates obtained by projecting 3D space time volume into a 2D representation. Polana[2] computed a spatio-temporal motion magnitude template as the basis for activities recognition Moreover, trajectory shapes allows to encode information about local motion. A hierarchical approach based on SIFT feature trajectories is proposed by [3]. Also in [4], feature trajectories are built by detecting and tracking spatial interest points. While in [5], the authors calculate the trajectories of detected 3D points using KLT tracker.

Feature extraction from video frames takes into account the temporal dimension. This is usually done by using optical flow vectors [6], spatial and temporal features such as cuboids [7].The selection of salient points is based on separable linear filters then cuboids are defined around these points [8].

Recently local descriptors have drawn much attention as a representation method for video content. They are able to capture motion and appearance. They are also robust to scale variation and the change of viewpoint. In addition, local features concept is extended to space-time domain. Therefore, most of the existing proposals consider the video as a space time volume and the local regions around salient points are described in 3D space. Space–time interest points are those points where the local neighborhood has a significant variation in both the spatial and the temporal domain. Yet, several spatiotemporal local feature descriptors and detectors have been proposed and evaluated in action recognition. Laptev [9] extracted Histogram of Gradient (HoG) and Histogram of Flow (HoF) from detected cuboids. [10] Propose a novel spatio-temporal feature employing SURF features and Lucas-Kanade optical flow detection methods. Also on [11] the SURF image descriptor is extended to a video descriptor called extended SURF (ESURF). In [12] Feature points are detected by the Spatio-Temporal Interest Points STIP algorithm then 3D patches are described around selected points. . In 3D local features point description approach, points are computed at random locations. While in 2D approaches, gradients are computed in polar coordinates. SIFT 3D [13] can be seen as an example of those approaches, the spatio-temporal gradients computation leads to problems due to singularities at the poles and induces progressively smaller bins at poles and induces progressively smaller bins at poles. Klaser [8] proposed an idea to resolve this problem by using regular polyhedrons and spherical coordinates to quantize the orientation of spatio-temporal gradient but with a high computational complexity. Laptev and Lindeberg [14] extended the Harris corner detector [15] to 3D while Klaser [16] combined Harris 3D detector with HOG 3D descriptor. Indeed, the representation of landmarks in 3D requires a preprocessing step and high operational complexity [11]. A weakness of such approaches is that it is difficult and time consuming to extend these features to the time domain. Also, descriptors are extremely high dimensional and they often retain redundant information.

Although much progress has been made in recent years there is still a conspicuous lack of descriptors with describe both motion and spatial information from video with a good compromise between average of recognition and computational complexity.

In this work we propose a new optimized approach for moving point's detection and description in video. This method is based on in 2D representation of the source video data called Accordion representation [17][18].

The rest of this paper is organized as follows. In section 2 we introduce the moving point detection and description process. Further, Evaluation framework is presented in section 3. The experiments and results discussions are presented in section 4 and section 5 concludes the work.

II. MOVING POINTS DETECTION AND DESCRIPTION

A. *Moving points detecion*

To extract moving points we have been combine tow wide technique on computer vision: background subtraction and optical flow algorithm. A graphical overview of our method is illustrated in figure 1.

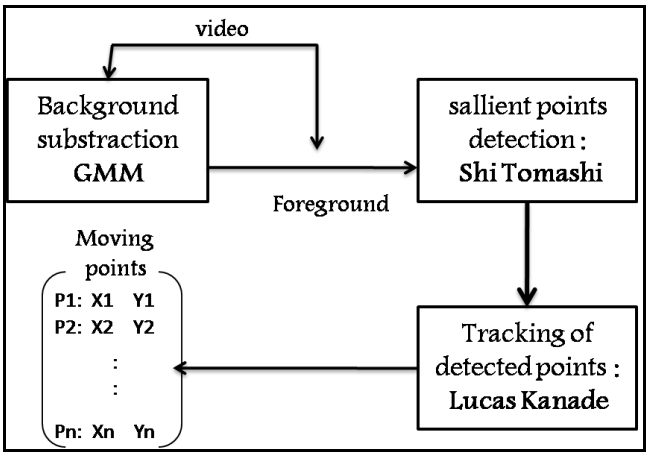


Figure 1. Moving points extraction process

In the first Step, background subtraction is applied on video to extract foreground points from background clutter in. We choose to use an adaptatif Gaussian Mixture Model (GMM) to detect motion. This technique is simple but the performances are very widely satisfactory when used in an environment without constraints. In addition, it has the advantage of requiring fewer resources in terms of computation and low memory space.

Given the large number of pixels in a video sequence, it would be far too heavy all trying to follow, this is why it is first necessary to select points that would be interesting to follow. We call point, a pixel window centered on a given pixel. Then, we use the Shi-Tomasi detector [19] to detect salient points in video frames. It is an extension of Harris detector. In fact, The Shi-Tomasi detector reduces the computational complexity of point's detection and explicitly chooses the points to be tracked more effectively. In the second step, detected salient points are tacked throughout the video frames by Lucas Kanade optical flow algorithm [20] to build trajectories.

B. *Accordion transformation*

As it was proved on [21], space-time frames contain motion information because there are formed by the temporal axe and the spatial one x or y, which makes them useful to describe action without the need to 3D extensions for the descriptors.

The Accordion transformation is obtained by carrying out a temporal decomposition of video. In fact, the temporal decomposition is done decomposing 3D volume video on a set of temporal images formed by the y axis and the time axis. To construct the image Accordion (IACC), temporal images are returned horizontally (mirror effect) as illustrated in figure 2. The last step consists in projecting these images successively on a 2D plane. [18] This representation can be obtained by traversing the temporal images while reversing the direction of movement of an image to another.

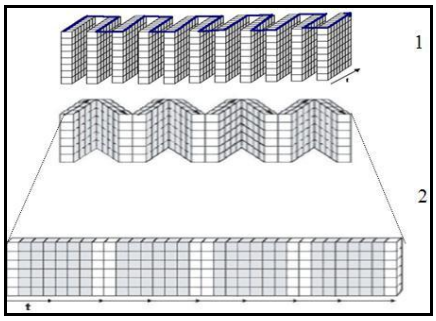


Figure 2. Accordion transformation

The Accordion image size $(Xacc, Yacc)$ is given by:

$$\begin{bmatrix} X = X \\ Y = Y * NF \end{bmatrix} \quad (1)$$

With X and Y are the video frames dimensions

The Accordion transformation tends to put in priority the exploitation of temporal correlation with transforming temporal correlation of pixels on the 3D video into high spatial correlation in the resulting 2D image (IACC).

C. *Motion description*

To describe motion information in a video sequence we need to detect moving points from the Accordion image. So, we have to compute the new position of each moving point in the 2D space of the IACC. The new Accordion coordinates of each point p(x,y) from the frame i are computed using the following projection function

Projection: video3D ->image2D
(x; y; i) ->(xAcc; yAcc)

We tested the moving point's detection process on tow videos "**v_jumping_10_01" and "v_juggle_05_01"** tacked from the YouTube action dataset. Figure 3 shows results of moving points detection.

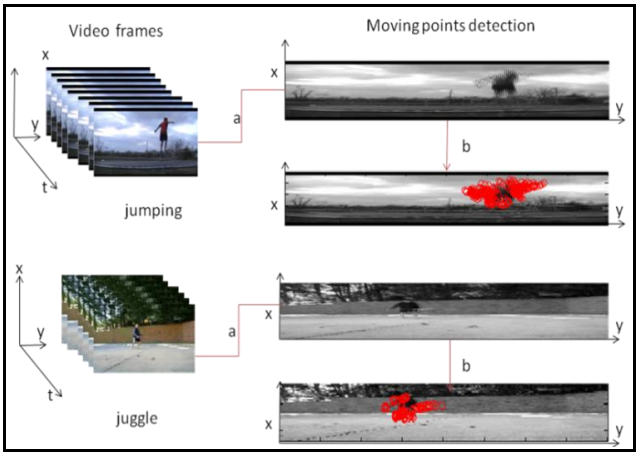


Figure 3. Moving points detection on IACC

In those examples the camera is fixed and the background is practically stable. As shown on the figure 3 each video is transformed into an image Accordion (a). Moving points extracted are projected onto this Image (b). Moving points are illustrated by a set of red points detected on IACC. We can conclude that the accordion transformation makes moving points in the same vicinity. So we have to exploit this spatial adjacency by calculating gradient directions between those pixels. Many suggested descriptors have proven to establish very good performance on local points detection and description such as Scale-Invariant Feature Transform (SIFT) [22] descriptor, Speeded Up Robust Features (SURF) [23], and Histogram of Oriented Gradients (HOG) [24]. All those descriptors applied locally to the IACC can capture the distribution of gradients within the moving point.

Making the moving points in the same neighborhood allows extracting the local motion descriptor around every point. In this works the motion vectors are computed using the local SIFT descriptor. SIFT describes the gradient distribution in the local neighborhood of a point of interest, it is robust to scale variation, rotation and it is claimed to be highly distinctive for discrimination. The size of the SIFT descriptor is equal to 128.

D. Spatial description of moving points

The video frames formed by the axes x and y are called spatial frames. We project moving points obtained in the previous section on M chosen frames from every N video frames. We apply the SIFT descriptor to extract appearance information of moving points. This is what we called spatial motion description in our method.

III. EVALUATION FRAMEWORK

Our evaluation framework is as follows. In the first step, moving point are extracted from each video. In the second step each video is transformed in IACC. Then, moving points are projected in the obtained IACC. For each moving point, local descriptor is computed. To make final videos signatures we investigate the bag of words approach. In fact, all computed motion descriptors for all moving points are quantized using the k-means clustering algorithm. We obtain a codebook of visual words. Then, each video is represented as a histogram of occurrence of the codebook elements. Finally, we use SVM for video classification.

In this set of experiments we evaluate the motion and spatial description of moving points. The datasets used are the KTH and Weizmann with the same experimental setup described in the previous sections.

Figure 4. Weizmann dataset

The Weizmann dataset contains 93 video sequences with a homogeneous and static background. It consists in ten types of action classes: *bending downwards, running, walking, skipping, jumping-jack, jumping forward, jumping in place, galloping sideways, waving with two hands, and waving with one hand*. Each action class is performed once (sometimes twice) by 9 subjects.

Figure 5. KTH dataset

The KTH data set contains six types of human *actions (boxing, hand-waving, handclapping, jogging, running and walking)* Those actions are performed several times by 25 subjects in different scenarios of outdoor and indoor environment with scale variations with scale change, with different clothes and indoors. Action classification is usually done by extracting descriptors from a training subset and comparing them to descriptors extracted from the testing videos

1) Motion description

In our experiments we vary the codebook size from 50 to 1100 for both Weizmann and KTH datasets. The variation of performance is given by the figure 12.The best recognition average is obtained with vocabulary size equal to 1000 for the tow datasets. We attempts respectively 92,4% and 85% for the Weizmann and the KHT datasets. Figures 6 and 7 present the variation of recognition rate.

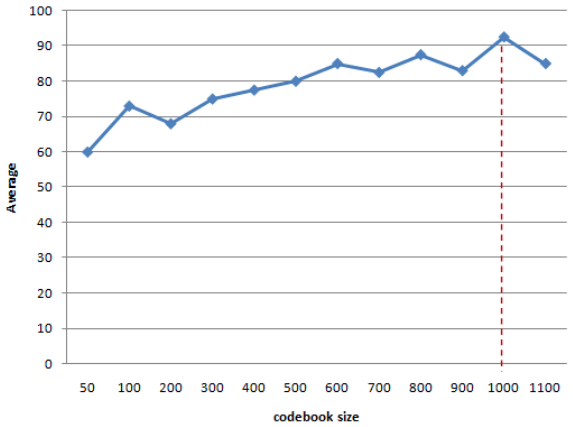


Figure 6 Recognition rate depending on the number of clusters for the Weizmann dataset

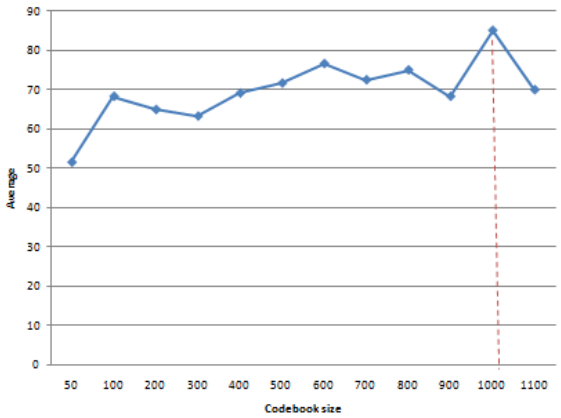


Figure 7. Recognition rate depending on the number of clusters for the KTH dataset

2) Spatial description of moving points

We extract SIFT descriptors from Moving points projected on some spatial frames of our video. We select N frames from every M frames. In the case of the Weizmann data set we choose M=3 and N=1, for the KTH data set we fixe M=10 and N= 2. We generate a vocabulary from those descriptors and we make signature of every video. We perform test by varying vocabulary size from 50 to 900 and 50 to 800 for respectively Weizmann and KTH dataset. The variation of recognition rate is plotted in figure 8 and figure 9. Best precisions are respectively 90,3% and 71,67% for the Weizmann and the KHT datasets

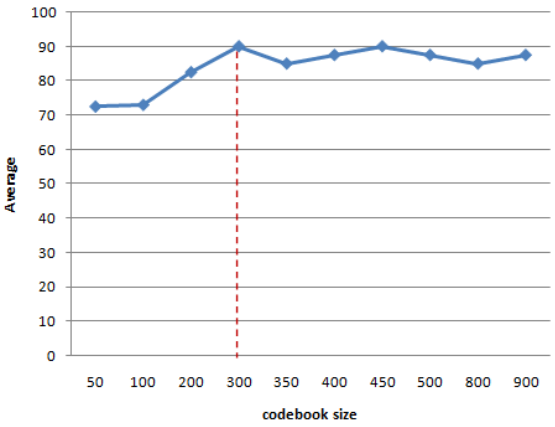
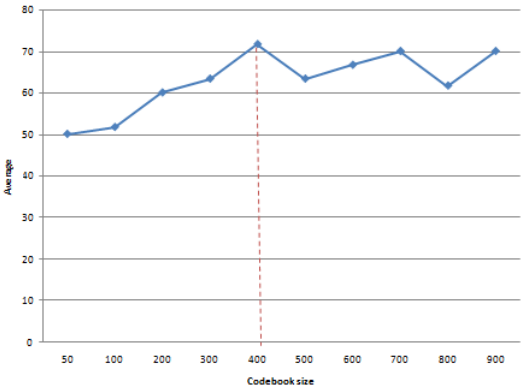


Figure 8. Recognition rate depending on the number of clusters for the Weizmann dataset



2 Figure 9. Recognition rate depending on the number of clusters for the Weizmann dataset

3) *Result Spatial-Motion description*

The result space-motion descriptor is built by concatenating the two preview descriptors. We obtain an average by 93,75 %. The M-SIFT addition improves accuracy of the spatial one. Table 1 shows the Categorization rates of the proposed descriptors on the two datasets.

TABLE I. ACCURACY OF PROPOSED DESCRIPTORS

Accuracy	**Descriptors**		
	Spatial-SIFT	***Motion-SIFT***	***M-S-SIFT***
KTH	71,67%	85%	78%
Weizmann	90,3%	92,4%	93,75%

These results provide a strong indication that 2D interest points descriptors can indeed be used to capture motion information in videos when applied to the IACC. The description of moving points on the spatial frames improve recognition rate when combing with the temporal description illustrated by the Motion-Spatial-SIFT descriptor.

The classification rates on KTH are less than the classification rate on Weizmann dataset. Indeed, the cameras used to collect the KTH videos are of low quality, and a high level of noise. The outdoors videos were captured by a hand-held camera, so there is motion in the background in most of the videos.

IV. EVALUATION

To evaluate our approach, we compare it to some previous works. In table 2 we report a comparison of the classification average of our descriptors with the state of the art reported by other works.

TABLE II. ACCURACY OF PROPOSED DESCRIPTORS

Method	**Accuracy KTH**	**Accuracy Wiezmann**
[16]	-	90,7%
[25]	81.5%	90,0%
[27]	-	90,4%
[8]	91.4%	84,3%
[6]	_	82,6%
[26]	_	72,8%
[28]	71,7%	_
[11]	84 ,5%	
Prposed S-SIFT	71,67%	90%
Proposed M-SIFT	85%	92,4%
Proposed M-S-SIFT	78%	93,75%

Result obtained on Weizmann dataset outperform the works proposed on [8] [6] [26]. In [6] points are seen as cuboids on 3D volume the space time descriptors proposed are based on the concept of HOG extended to 3D space . Also the results reported by Liu [27] (90.4%). Our results outperform also those of Klaser [16] with a heavy optimization of descriptor size and computational complexity even those based on SIFT descriptor like Lopez and also the results reported by SIFT3D. The SIFT 3D descriptor is 4096 dimensional descriptor however we can release much better average with descriptor sizes equal to 128 for both Motion SIFT and spatial SIFT and equal to 256 for M-S-SIFT.

Also when considering the KTH dataset we outperform previous BoW based works [28, 11].

Also, we tested the time complexity for the two proposed descriptors for the KTH data set, as illustrated on table 3, all times are much reduced and descriptors are extracted with a very small moving points number.

TABLE III. TIME COMPLEXITY COMPARISON FOR KTH DATASET

Descriptor	**M-SIFT-ACC**	**S-SIFT**
Detection and Description	62s	62s
Moving points detection	7.56 s	7.56 s
Histogram computation	217 s	66 s
Moving points number	816	816

In table 4, we presents the time complexity calculated for the Weizmann data set and we compare it with some results founded on the state of the art on the same dataset.

TABLE IV. TIME COMPLEXITY COMPARISON FOR WEIZMANN DATASET

Descriptor	STIP[15]	ST-SIFT[22]	M-SIFT-ACC	S-SIFT
Detection And Description	582s	1327 s	27.33 s	20,12s
Moving points detection	_____	_____	3.28 s	3.28s
Histogram computation	113s	1395 s	104 s	38 s
Moving points number	10886	504766	214	214

We can see on table 4 that the detection and description time is less than that of STIP which is a 3D salient point's detector and descriptor.

Despite the 2D video representation for the ST-SIFT we reached more optimized time detection and description and a more reduced histogram computation time. The number of detected moving points is also reduced.

V. CONCLUSION

Extending features to 3D is the predominant methodology in video action recognition. The biggest description challenge is the fact that observed video appearances for each human action contains large variances in body poses, non-rigid body movements and clothing texture. This is why we choose to build an hybrid descriptor composed of both static features (local appearance) and motion features (local motion) to develop an effective recognition framework based of the bag of words approach. In this work we show that using the 2D accordion representation is adequate for detection of spatio-temporal feature. Despite the simplicity of our approach, we achieved good result with reduced size descriptors, lower computation complexity and small codebook sizes.

VI. ACKNOWLEDGMENT

The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB

VII. REFERENCES

[1] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates", *IEEE T-PAMI*, pp 257-267, 2001

[2] Polana, R., and Nelson, R. 1994, Low level recognition of human motion (or how to get your man without finding his body parts). In Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects, p. 77-82, Austin TX, 1994

[3] 2. J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, andJ. Li. Hierarchical spatio-temporal context modeling for action Recognition. In CVPR, 2009.3. .

[4] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In ICCV workshop on Video-oriented Object and Event Classification, 2009.

[5] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In ICCV, 2009.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features, In VS-PETS'05, 2005

[7] S.Ali, and M.Shah, "Human action recognition in videos using kinematic features and multiple instance learning", in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2010

[8] A. Klaser, M. Marsza lek, and C. Schmid, "A spatio-temporal descriptor based on 3Dgradients", In BMVC'08, 2008

[9] Laptev, M. Marsza lek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies", In CVPR'08, 2008

[10] A SURF-based Spatio-Temporal Feature for Feature-fusion-based Action Recognition, Akitsugu Noguchi, Keiji Yanai, Third Workshop on HUMAN MOTION Understanding, Modeling,

[11] G. Willems, T. Tuytelaars, and L. Van Gool, "An effcient dense and scale-invariant spatio-temporal interest point detector", In ECCV'08, 2008

[12] I. Laptev and T. Lindeberg, "Space-time interest points", In ICCV, 2003

[13] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition", In MULTIMEDIA, 2007

[14] Laptev, I., Lindeberg, T.: Space-time interest points. In: Proc. IEEE Intl. Conf.on Computer Vision. Volume 2. (2003) 432–439

[15] Harris, C., Stephens, M.: A combined corner and edge detection. In: Proc. Alvey Vision Conf. (1988) 147–151 5. Schmid, C., Mohr, R.: Local grayvalue

[16] A. Klaser, M. Marsza lek, C. Schmid, and A. Zisserman,"Human Focused Action Localization in Video", in International Workshop on Sign, Gesture, Activity 2010

[17] T.Ouni, W.Ayedi and M.Abid, " New low complexity DCT based video compression method", In Proceedings of the 16th International Conference on Telecommunications (ICT'09), 202-207, Piscataway, NJ, USA, 2009, IEEE Press

[18] T.Ouni, W.Ayedi and M.Abid, "New Non Predictive WaveletBased Video Coder: Performances Analysis", In Proceedings of International Conference on Image Analysis and Recognition., pp 344-353, Berlin, Heidelberg, 2010. Springer-VerlagIntelligence 2000, pp. 747-757

[19] C.Tomasi and T.Kanade, Detection and tracking of Point Features, Carnegie Mellon University TeChnical Report CMU-CS-91-132, April 1991

[20] J.Y Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm", Intel Corporation, Microprocessor Research Labs, 1999

[21] A.P.B.Lopes, R.S. Oliveira, J.M. de Almeida, and A.de Albuquerque Araujo, Spatio-temporal frames in a bag-of-visual-features approach for human actions recognition, in SIBGRAPI 09. IEEE Computer Society, 2009

[22] D. Lowe,"Distinctive image features from scale-invariant keypoints", IJCV'04,pp 91-110, 2004

[23] Bay H., Ess A., Tuytelaars T., Gool L.V.: SURF: Speeded Up Robust Features. In: Computer Vision and Image Understanding, 2008.

[24] 3. Dalal N., Triggs B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Conference on Computer Vision and Pattern Recognition, 2005

[25] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words", IJCV'08, pp 299-318, 2008

[26] J. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classiffcation", In CVPR'07, 2007

[27] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features", In CVPR'08, 2008

[28] C. Schuldt, I. Laptev, and B. Caputo. Recognizing humanactions: a local SVM approach. In Proc. of ICPR, 2004