Action Recognition form Videos

Sushirdeep Narayana
Department of Electrical and Computer Engineering
Texas A&M University
College Station, United States
sushir_369@tamu.edu

***Abstract*—In this project an action recognition system was developed using 3-dimensional SIFT descriptor and multiclass SVM classifier. The video trials were classified according to their specific actions based on the features extracted using a Bag-of-Features model. 3-D SIFT descriptors were employed on the Harris Laplace interest points of the video frames to extract spatio-temporal information. Visual Vocabulary was constructed to compute the features representing the Bag of Visual Words. The one vs. one and the one vs. all SVM multiclass classification were explored with this problem. The classifier models obtained from these techniques were evaluated.**

Keywords—SIFT; Multiclass SVM; Action Recogntion; Harris Corners; Bag-of-Features

I. INTRODUCTION (*HEADING 1*)

Human actions reveal primary descriptive information in videos. Action Recognition is the task of automatically understanding video data. In the past few years, the field of visual recognition had an outstanding evolution. Much of this progress has been sparked by the creation of realistic datasets as well as by the new robust methods for image description and classification. Classification of videos based on different actions performed is a very challenging problem. To solve this problem concepts from Machine Learning and Computer Vision must be applied. The applications of human activity recognition include video surveillance, video information retrieval, content-based browsing, video recycling and human-computer interaction. Recognition of human actions from videos is a difficult task. We must deal with significant intra-class variations, background clutter, and occlusions.

Space-time interest points have been proposed to capture local events in the video frame. Such points have stable locations in space-time and provide a potential basis for part-based representations of complex motions in a video. The central theme of this project is the recognition of Spatio-temporal events and activities by applying multiclass Support Vector Machines (SVM) to the features obtained from the Bag of Visual Words Model. The Bag of Visual Words is constructed using local Spatio-temporal descriptors from 3D SIFT with the assistance of Harris-Laplace interest points. The approach can hence be an extension of previous interest point based spatial recognition methods into space-time.

Section 2 describes some of the related work in the domain of action recognition from videos. Section 3 explores the Action Classification methodology. Section 4 reports the Experimental Setup and the Results are given in Section 5. Some significant inferences and conclusions are drawn in Section 6.

II. RELATED WORK

There has been quite a lot of work in Action Recognition from videos. In [2] the authors use motion characteristics to capture the relevant features from a video. Two types of motion characteristics are explicitly added as features, the dominant motion, and the residual motion. The authors use Divergence, Curl and Shear (DCS) motion descriptor to represent the video. [4] use Speed Up Robust Feature (SURF) descriptors and optical flow to capture the spatial and temporal characteristics of a video sequence. In [8], an "action bank" approach was proposed. The action bank representation is a concatenation of volumetric max-pooled detection volume features from each detector. This high-level representation transfers the semantics from the bank entries through to the output. The procedure used in the action bank representation is shown in Figure 1.

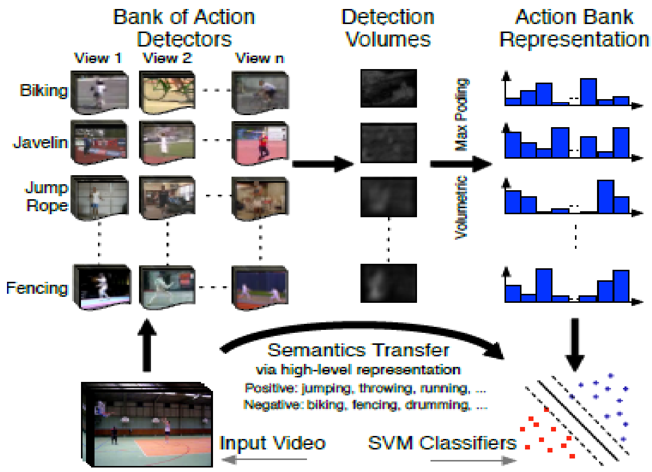


Figure 1: Action Bank Classification from Videos

In this project, a local Spatio-temporal descriptor, 3D Scale Invariant Feature Transform (SIFT) was used to extract the space-time descriptors because SIFT is known for being a robust and accurate descriptor.

III. Action Classification

Videos are three-dimensional variables. In this project, they are considered as a combination of frames over a period of time. The steps involved in the proposed action classification framework is given in this section. The first step is to apply Harris-Laplace corner detectors to identify the important spatial regions in the video. The next step is to describe the Spatio-temporal regions around these points using the [3] 3D SIFT descriptor. The descriptors are assembled from all the interest points for every fifteenth frame of the video to reduce computational time and increase the computational efficiency. The accumulated descriptors are quantized using k-means clustering and a Bag of Visual Words feature vectors are constructed. Finally, multiclass SVM is used to obtain the representative models for each action. These steps are shown as a flow diagram in Figure 2. In this project six human actions were recognized, namely, walking, jogging, running, boxing, handwaving, and handclapping.

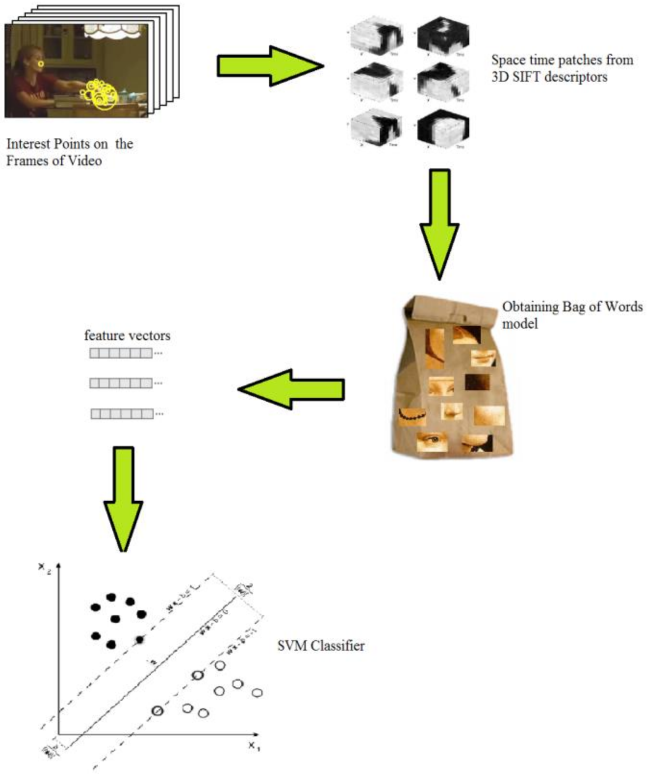

Interest Points on the Frames of Video
Space time patches from 3D SIFT descriptors
Obtaining Bag of Words model
feature vectors
SVM Classifier

Figure 2: Proposed Action Classification Approach

A. Harris-Laplace Corner Interest Points

The Harris-Laplace Corner Interest points [6] are robust to changes in scale, image rotation, illumination, and camera noise. In addition, they are highly discriminative. The method first builds up two separate scale spaces for the Harris function and the Laplacian. It then uses the Harris function to localize candidate points on each scale level and selects those points for which the Laplacian simultaneously attains an extremum over scales. To simplify, the fundamental property used is that shifting an image window in any direction around the corner should give a large change in intensity. The equation describing the window-averaged change of intensity is given below

$$E(u,v) = \sum_{x,y} w(x,y)[I(x+u,y+v) - I(x,y)]^2 \quad (1)$$

where *w(x,y)* is the window function , *I(x+u, y+v)* represents the shifted intensity of the image and *I(x,y)* represents the intensity of the image. Figure 3 illustrates the equation in a nice fashion. Figure 4 shows the application of Harris-Laplace Interest Points in an image frame of the video.

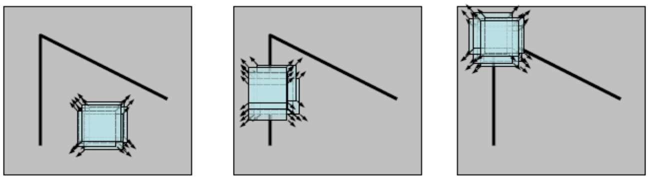


Figure 3: Harris Corner Detection

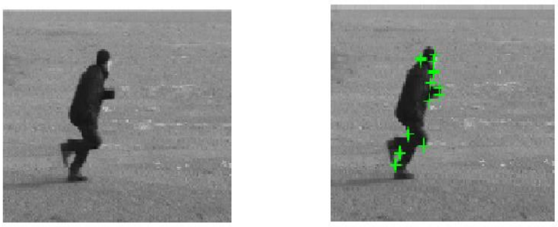


Figure 4: Harris-Laplace Interest Points on an image frame of the video

B. 3D SIFT Descriptor

Once the Harris-Laplace interest points are computed on the video frames, the 3D SIFT descriptor is used to describe the information in the Spatio-temporal space of the video in the region associated with the interest points computed.

3D SIFT descriptor [3] is an extension of the 2D SIFT descriptor in an intuitive manner. 3D SIFT descriptor [] is not 2D SIFT on a series of image frames of the video. There is a slight difference as depicted by Figure 5. The equations for computing the gradient magnitude and the orientations in the 3D SIFT descriptor are given below

$$m_{3D}(x,y,t) = \sqrt{L_x^2 + L_y^2 + L_t^2} \qquad (2)$$

$$\theta(x,y,t) = tan^{-1}\left(\frac{L_y}{L_x}\right) \qquad (3)$$

$$\varphi(x,y,t) = tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}\right) \qquad (4)$$

It is evident that φ encodes the angle away from the 2D gradient direction. Each pixel now has two values that denote the direction of the gradient in three dimensions.

The orientation assignment is achieved by taking each *(x,y,z)* position in the neighborhood and multiply it by the matrix given below

$$\begin{bmatrix} cos\theta cos\varphi & -sin\theta & -cos\theta sin\varphi \\ sin\theta cos\varphi & cos\theta & -sin\theta sin\varphi \\ sin\varphi & 0 & cos\varphi \end{bmatrix}. \quad (5)$$

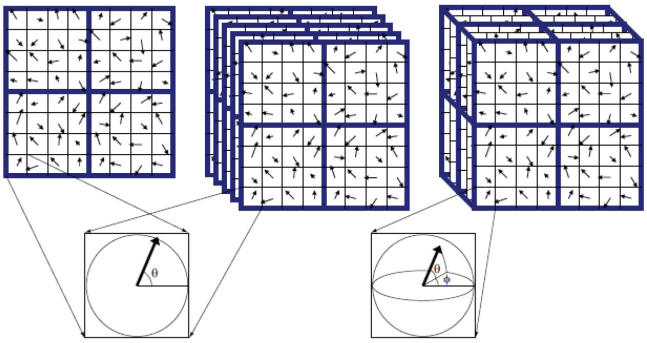


Figure 5: Image on the left shows a 2D SIFT descriptor on a series of image frames of the video and the image on the right shows a 3D SIFT [] descriptor with its 3D sub-volumes accumulated on the video

The length of the 3D SIFT descriptors is based on the number of bins used to represent θ and φ angles. The dimension of the 3D SIFT descriptor was chosen as 2160. In this project, 100 descriptors have been selected for each video, and each 3D SIFT descriptor was computed for every 15th frame of the video to reduce computational time and increase computational efficiency.

C. *Constructing Bag of Visual Words*

The bag of visual words [10] is a vector of occurrence counts of a vocabulary of the 3D SIFT features. To obtain the features as a Bag of Visual Words the following procedure was followed. First, the 3D SIFT descriptors are accumulated, and the k-means clustering is applied to quantize them to a specified number of clusters. The k-means algorithm iteratively groups the descriptors into k mutually exclusive clusters. The resulting clusters are compact and separated by similar characteristics. Each cluster center represents a feature or visual word. The collection of clusters is called "Spatio-temporal word vocabulary". This process can be visualized using Figure 6.

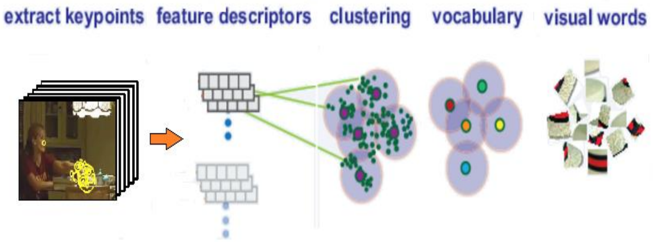


Figure 6: Constructing the Visual Vocabulary from the descriptors

The descriptors from the videos are matched to each cluster center, and the frequency of these cluster matches is converted to a frequency histogram. The procedure for obtaining the feature vectors is shown in Figure 7. Each feature represents the number of descriptors belonging to that cluster.

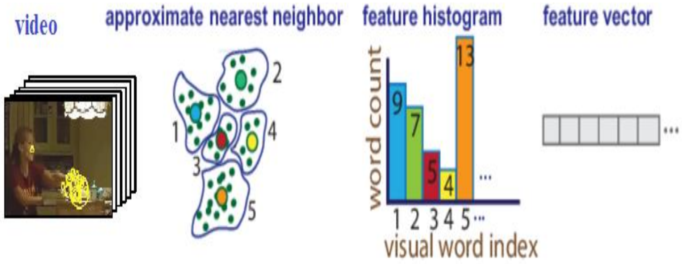


Figure 7: Computing the feature vectors

In this project the value of k =600 was selected. The feature vectors had a dimension of 1x k for each example, each dimension representing the co-occurrences of the descriptors with the cluster centers.

D. *Multiclass Support Vector Machines*

In this project, the one vs. one and the one vs. rest schemes of multiclass Support Vector Machines were explored. Let's assume there are K > 2 classes in the labeled data, where K denotes the number of classes. This type of problem is identified as Multiclass Classification. There are two well-known approaches to multiclass classification

- One vs. Rest: For each binary learner, one class is positive, and the rest are negative as shown in Figure 8. This design exhausts all combinations of positive class assignments.
- One vs. One: For each binary learner, one class is positive, another is negative, and the design ignores the rest as shown in Figure 9. This design exhausts all combinations of class pair assignments.

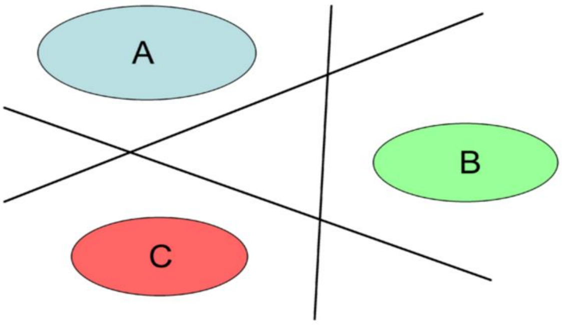


Figure 8: One vs. Rest Classification Scheme

For the rest of the section, the binary SVM will be discussed as Multiclass SVMs are an extension of the binary SVMs using the strategies mentioned above.

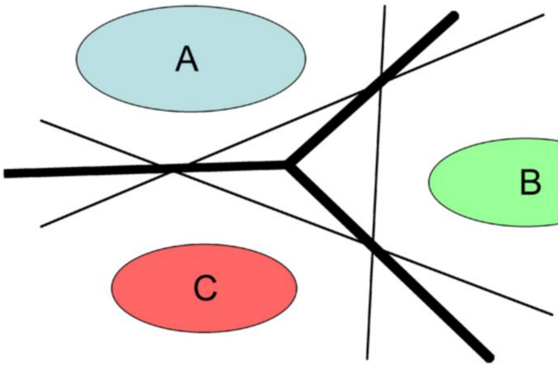


Figure 9: One vs One Classification scheme

The Support Vector Machine [11] approaches the problem of classification through the concept of the *margin*, which is defined as the smallest distance between the decision boundary and any of the samples as illustrated in Figure 10. In SVMs the decision boundary is chosen to be the one to which the margins are said to be maximized. The processes of maximizing the margin is described by the equations below

$$y(x) = w^T\varphi(x) + b \quad (6)$$

where *φ(x)* represents the feature space transformation of *x* and *b* denotes the bias parameter. Let t_n denote the binary class labels, then the maximizing the margin is depicted in the equation below

$$\arg max_{w,b}\left\{\frac{1}{||w||} min_n[b_n(w^T\ \varphi(x_n) + b)]\right\} \quad (7)$$

We choose,

$$t_n(w^T\varphi(x_n) + b) = 1 \quad (8),$$

for the point that is closest to the surface. In that case, the optimization problem reduces to

$arg_{w,b}\ min\frac{1}{2}\ ||w||^2$, such that,

$$t_n(w^T\varphi(x_n) + b) \geq 1\ \ , where\ , n = 1,2,\dots,m \quad (9)$$

where there are m training points. The above equation represents a Quadratic Programming Problem. In order to solve this constrained optimization problem, we use the Lagrange multipliers $a_n \geq 0$, and maximize the Lagrangian

$$\max L(w,b,a) = \tfrac{1}{2}||w||^2 - \textstyle\sum_n a_n\{t_n(w^T\varphi(x_n)_b - 1)\}\ . \quad (10)$$

Solving the above optimization problem, we get,

$$w = \sum_n a_n t_n \varphi(x_n) \quad (11)$$

$$0 = \sum_n a_n t_n \quad (12)$$

Eliminating w and b from L(w,b,a) gives the dual representation of the problem as

$$\max L^{\sim}(a) = \sum_n a_n - \frac{1}{2}\sum_n\sum_m a_n a_m t_n t_m\, k(x_n, x_m)$$

with respect to the constraints

$$a_n\ \geq 0\ and \sum_n a_n t_n = 0 \quad (13)$$

The classification now occurs as per the equations below

$$y(x) = \sum_n a_n t_n k(x, x_n) + b$$

such that the Karush-Kuhn-Tucker (KKT) conditions are satisfied as given below

$$a_n \geq 0\ \ ,$$

$$t_n y(x_n) - 1 \geq 0\ and$$

$$a_n\{t_n y(x_n) - 1\} = 0 \quad (14)$$

In this project, linear kernel was used and the box constraint C parameter were chosen using cross validation.

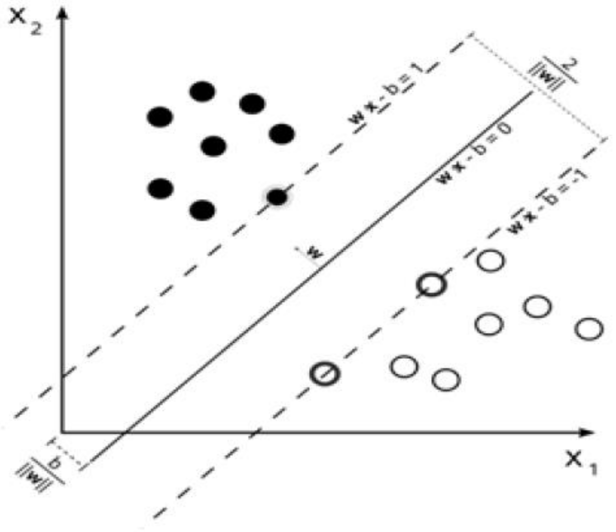


Figure 10: Illustration of Maximum Margin Hyperplane in SVMs

IV. Experimental Setup

The database that was used is the KTH [1] Action Database. The database contains approximately around 600 videos. Sample image frames of the some of the videos are shown in Figure 11. There were six classes of human action in the videos, namely, jogging, walking, running, boxing, handwaving, and handclapping. 25fps was the frame rate of the videos. The resolution of the video frames is 160x120. Programs were composed in Matlab 2016a. The dataset was partitioned such that 80% belonged to the training set and 20% belonged to the testing set. The box constraint C parameter was chosen using 5-fold Cross Validation and the prediction results obtained are reported for the selected C.

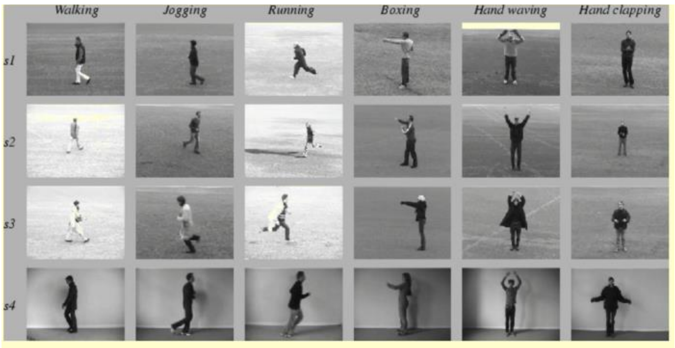
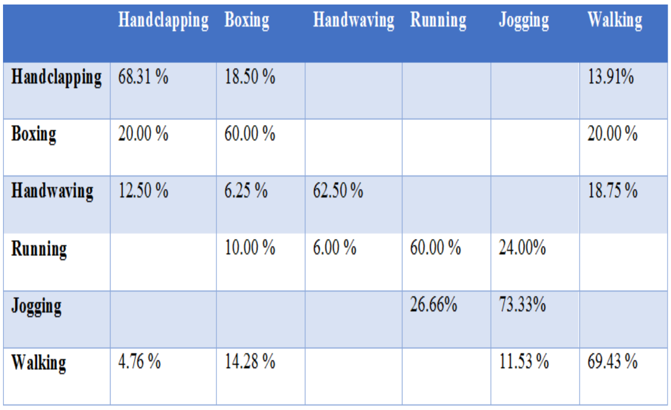


Figure 11: KTH Video Database

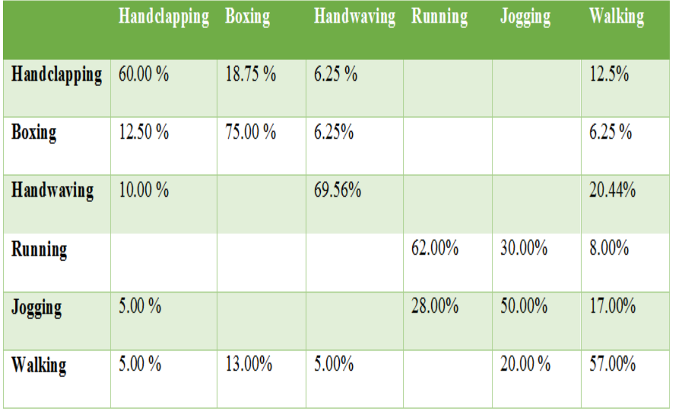
V. Results

The primary goal of this project was to classify human actions using the KTH Video Dataset. The dimensions of the 3D SIFT Descriptors were chosen as 2160. Descriptor dimension of 640 was also tried but gave bad results. The k-means clustering was performed with k=600. The one vs. rest and the one vs. one Multiclass SVM classification were performed. The test error for one vs. one Multiclass SVM was obtained as 65.595 %. The confusion matrix computed by using one vs. one is illustrated in Figure 12. The test error for one vs. rest Multiclass SVM was achieved as 62.26 %. The confusion matrix for this is given in Figure 13. Each box in the Figures 12 and 13 denote accuracy.



	Handclapping	Boxing	Handwaving	Running	Jogging	Walking
Handclapping	68.31 %	18.50 %				13.91%
Boxing	20.00 %	60.00 %				20.00 %
Handwaving	12.50 %	6.25 %	62.50 %			18.75 %
Running		10.00 %	6.00 %	60.00 %	24.00%	
Jogging				26.66%	73.33%	
Walking	4.76 %	14.28 %			11.53 %	69.43 %

Figure 12: One vs One Multiclass SVM Confusion Matrix



	Handclapping	Boxing	Handwaving	Running	Jogging	Walking
Handclapping	60.00 %	18.75 %	6.25 %			12.5%
Boxing	12.50 %	75.00 %	6.25%			6.25 %
Handwaving	10.00 %		69.56%			20.44%
Running				62.00%	30.00%	8.00%
Jogging	5.00 %			28.00%	50.00%	17.00%
Walking	5.00 %	13.00%	5.00%		20.00 %	57.00%

Figure 13: One vs Rest Multiclass SVM Confusion Matrix

VI. Conclusion and Discussion

In this project, the videos were classified as per their human actions. There are many elements involved in this process. The one vs. one Multiclass SVM performed better than the one vs. rest scheme of Multiclass SVM. The descriptors were computed only once in every 15th frame. Only 100 descriptors for each video were used to construct the Visual Vocabulary. This was done for less computation and saving computation time. Increasing the number of descriptors per video would have improved the accuracy. Also, increasing the dimension of the 3D SIFT descriptor further might help in improving the accuracy. The k parameter of the k-means clustering is another important parameter that can impact the classification performance. The value of k = 600 gave better performance results compared to k=1000 and k =2000 that were tried. SVMs with linear kernels gave better performance than SVM with Gaussian kernels for this problem. The computation time involved was very high as the videos are 3-dimensional. Applying Parallel Computing would be one way to speed up computation and gather more features.

References

[1] C. Schüldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach”, In ICPR, 2004.

[2] I. Laptev, “On time-space interest points”, IJCV, 64(2/3): 107-123,2005.

[3] P Scovanner, S Ali, M Shah, “A 3-dimensional sift descriptor and its application to action recognition”, Proceedings of the 15th ACM international conference on Multimedia, 357-360B.

[4] Wang, A. Klaser, C. Schmid, and C. L. Liu, “Action Recogntion by dense trajectories”, In CVPR, 2011.

[5] Solmaz, S. Assari and M. Shah, “ Classifying web videos using a global video descriptor”, MVAP-d-12-00244.

[6] C. Harris and M.J. Stephens, “A combined corner and edge detector”, In Alvey Vision Conference, 1988.

[7] Ivan Laptev, "Local Spatio-Temporal Image Features for Motion Interpretation", PhD Thesis, 2004, Computational Vision and Active Perception Laboratory (CVAP), NADA, KTH, Stockholm.

[8] D.G. Lowe, “Distinctive Image Features from Scale Invarient Keypoints”, in International Journal of Computer Vision (IJCV), 2004.

[9] S. Sadanand, J. Coroso, “Action Bank: A High Level Representation of Activity in Video”, CVPR, 2012.

[10] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints”, In ECCV Workshop on Statistical Learning in Computer Vision, 2004.

[11] Christopher Bishop, “Pattern Recogntion and Machine Learning”, Springer, 2007

[12] R. Szeliski, “Computer Vision: Algorithms and Applications”, Springer