Video-Based Human Action Recognition Report

ABSTRACT: In this paper, video-based action recognition is performed on KTH dataset using four combinations of two feature descriptors Histogram of Oriented Gradient Descriptor (HOG) and 3-dimensional Scale Invariant Feature Transform (3D SIFT) and two classifiers Support Vector Machine (SVM) classifier and K Nearest Neighbour (KNN) classifier. Features are extracted from extracted frames of training videos using descriptor and clustered to form Bag-of-words model. This model is used to construct histograms which are given to the classifier to create a training model. Test videos are classified into handclapping, running, boxing, waving, walking and jogging actions using trained model of classifier. Accuracies of action recognition obtained for four combinations: HOG-SVM, HOG-KNN, 3D SIFT-SVM and 3D SIFT-KNN are compared to find the best combination for action recognition. Also, the effect of k parameter of k-means used to generate Bag-of-words model is evaluated for action recognition.

KEYWORDS: Computer Vision, Human Action Recognition, HOG Descriptor, 3D SIFT Descriptor, SVM Classifier, KNN Classifier, STIP, Bag-of-Words model.

I. INTRODUCTION

Computer Vision is a field of Computer Science and Artificial Intelligence that obtains information from images and videos and helps in better understanding of 2-dimentional and 3-dimentional images. Computer vision deals with interpreting and obtaining meaningful description of objects from their images or videos. Computer vision has variety of applications in fields like medicine, robotics, security, transportation and industrial automation.

Action recognition is an important problem in computer vision. It finds applications in surveillance systems for security, search engines for image identification, detection of abandoned object, human vehicle and human computer interactions, video analysis for detection of abnormal or illegal activities, traffic monitoring and healthcare monitoring for patient [1]. It is therefore necessary to use efficient techniques for action recognition in order to obtain accurate information required for such applications [2]. The efficient choice of technique will save time as well as cost as most accurate solution will be available in less time for a particular application.

Conventional approach for human action recognition includes optical flow computation [3]. It has a disadvantage of failing to capture sudden changes in motion. In constantly changing image structures in videos certain points with non-constant motion may contain important information regarding change in structure. Traditional human recognition techniques were based on local feature extraction algorithms on 2D images like 2D SIFT [4] descriptors. However, 2D images cannot retain the 3D information and are susceptible to damage due to external conditions. 3D SIFT [5] descriptors are proposed to overcome the challenges faced by 2D SIFT.

The dataset used is KTH dataset. It consists of 6 types of human actions that is running, walking, handclapping, handwaving, boxing and jogging performed by 25 subjects in four different scenarios [6].

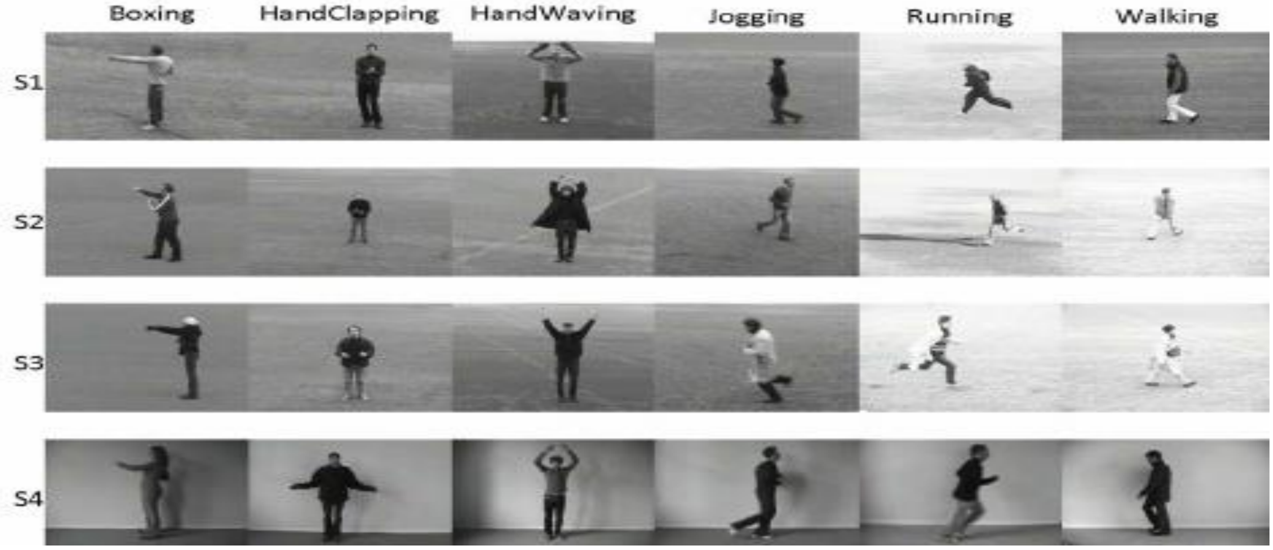

Fig. 1. KTH dataset with six actions with 4 scenarios [6]

The paper compares the efficiency of HOG and 3D SIFT feature descriptors used with SVM and KNN classifiers for human action recognition from videos. Thus, there are four combinations compared for human action recognition efficiency that is HOG and SVM, HOG and KNN, 3D SIFT and SVM and 3D SIFT and KNN. This paper is organized as follows: Feature descriptors and classifiers used are explained in section II. Methodology is explained in section III. Results are described in section IV. Conclusion of the work is provided in section V.

II. **BACKGROUND STUDY**

Histogram of Oriented Gradients [7] is a feature descriptor used for the detection of objects like people, vehicles, trees, etc. HOG is extended from 2D to 3D by including the temporal information [8].

Scale Invariant Feature Transform [4] is a feature descriptor used in computer vision to describe local features. 3D SIFT is an extension to SIFT used in video based action recognition. The equations for computing the gradient magnitude and the orientations in the 3D SIFT descriptor are given below:

$$m_{3D}(x,y,t) = \sqrt{L_x^2 + L_y^2 + L_z^2} \qquad (1)$$

$$\theta(x,y,t) = \tan^{-1}\left(\frac{L_y}{L_x}\right) \qquad (2)$$

$$\varphi(x,y,t) = \tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}\right) \qquad (3)$$

The direction of gradient is computed in three dimensions for each pixel. For orientation assignment the position value (x,y,t) in the neighborhood is multiplied by the following matrix:

$$\begin{bmatrix} \cos\theta\cos\varphi & -\sin\theta & -\cos\theta\sin\varphi \\ \sin\theta\cos\varphi & \cos\theta & -\sin\theta\sin\varphi \\ \sin\varphi & 0 & \cos\varphi \end{bmatrix} \qquad (4)$$

In this paper, 50 descriptors of dimension 640 have been selected for each video. These descriptors are then clustered to form spatio-temporal Bag of Words model [9]. A bag of keypoints represents a histogram of the number of occurrences of particular pattern in a given image [9]. It helps to improve the computation efficiency of the classifier.

Support vector machines [10] are a type of supervised learning algorithms. A support vector machine is a linear classifier based on the concept of a separating hyperplane.

SVMs use linear models for two class classification problems given by following equation:

$$y(x) = w^T\,\Phi(x) + b \qquad (5)$$

where b is the bias and $\Phi(x)$ represents a fixed feature-space transformation [10].

To solve multiclass problems, we have to combine many two class SVMs to develop a multiclass classifier.

KNN [11] is a lazy learning classification algorithm that uses similarity measure to classify new classes using stored classes. A Euclidean distance function is one of the distance functions used to measure the distance between training samples and the test input [11]. Euclidean distance is defined as:

$$d(i,j) = \sqrt{\left(\sum_{i=0}^{n}(x_i - y_i)^2\right)} \qquad (6)$$

Spatio-temporal features [12] are located at spatio-temporal salient points that are extracted with interest point operators. Space time interest point detectors are extensions of 2D interest point detectors that include temporal information along with spatial information. Spatio-temporal Interest points (STIP) were proposed by I. Laptev in 2005 [12].

K-means [13] is unsupervised clustering algorithm used to solve clustering problem. The distance measures include the Euclidean, Manhattan and Minkowski distance [13].

Algorithm steps are as follows:

1. Initialize the center of the clusters
2. Attribute the closest cluster to each data point
3. Set the position of each cluster to the mean of all data points belonging to that cluster
4. Repeat steps 2-3 until no change

III. PROPOSED METHODOLOGY

KTH dataset videos are used for training and testing with six action classes boxing, handclapping, running, handwaving, jogging and walking. Out of 100 videos of each class 80 videos are used for training and 20 videos are used for testing purpose. The classifiers are trained on the training set while recognition results were obtained on the test set as well as training set.

A. Histogram of Oriented Gradient Methodology

1. Computation of Space Time Interest Points

   Initially space time interest points (STIPs) are computed for each video. Then frames are extracted from each video and resized into 160x120. These are then used by HOG to calculate the descriptor.

2. Extraction of Features

   HOG features are extracted for each video using STIP points and frames. Features can be used for object detection, tracking and classification.

3. Creation of Bag-of-Words model

   The k-means clustering algorithm is used to create the bag-of-words model from the descriptors of training videos. Histograms are constructed for each video using HOG descriptors and centroid clusters generated after k-means.

4. Training the classifier

   The histograms are used to train the classifier. The classifier will then generate the trained model. This model is used for classifying the test videos.

5. Classification of Test videos

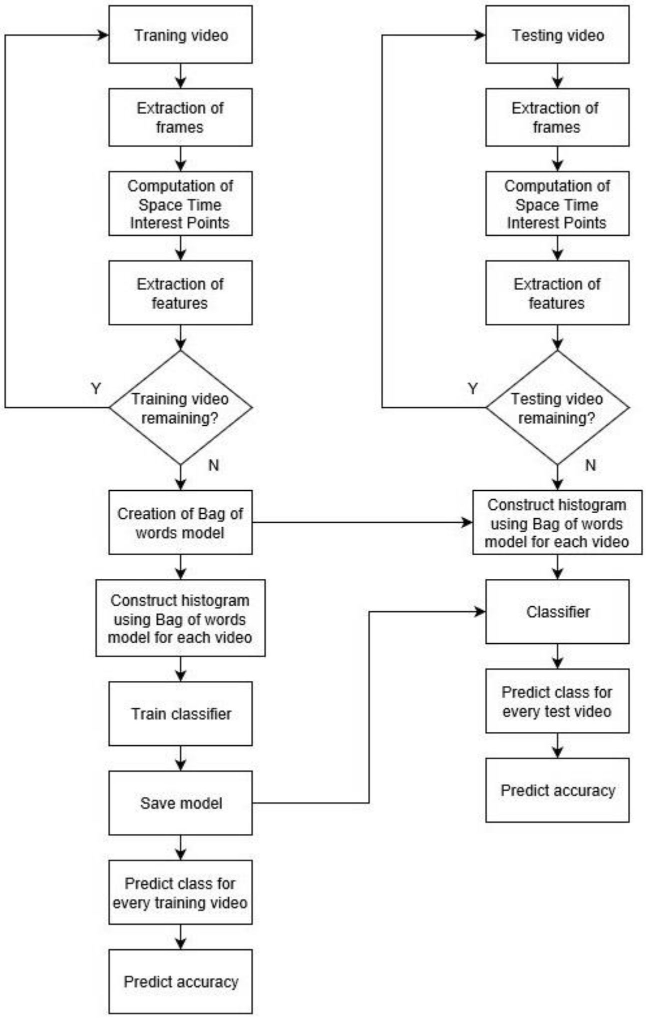
Traning video
Extraction of frames
Computation of Space Time Interest Points
Extraction of features
Training video remaining?
Y
N
Creation of Bag of words model
Construct histogram using Bag of words model for each video
Train classifier
Save model
Predict class for every training video
Predict accuracy
Testing video
Extraction of frames
Computation of Space Time Interest Points
Extraction of features
Testing video remaining?
Y
N
Construct histogram using Bag of words model for each video
Classifier
Predict class for every test video
Predict accuracy

Fig. 2. Methodology for video based human action recognition

Histograms of test videos are given to the classifier. The classifier uses the trained model to label the videos. Accuracy is computed for test videos.

A. 3D Scale Invariant Feature Transform Methodology

1. Computation of Space Time Interest Points

Initially space time interest points (STIPs) are computed for each video. Then frames are extracted from each video and resized into 160x120. These are then used by 3D SIFT to calculate the descriptor.

2. Extraction of Features

3D SIFT features are extracted from each video using STIP and extracted frames. These features are invariant to rotation and scale.

3. Creation of Bag-of-Words model

The k-means clustering algorithm is used to create the bag-of-words model from the descriptors of training videos. Histograms are constructed for each video using 3D SIFT descriptors and centroid clusters generated after k-means.

4. Training the classifier

The histograms are used to train the classifier. The classifier will then generate the trained model. This model is used for classifying the test videos.

5. Classification of Test videos

Histograms of test videos are given to the classifier. The classifier uses the trained model to label the test videos. Accuracy is computed for test videos.

IV. Results and Discussion

The primary goal of the paper is to compare the classification results obtained for four combinations: HOG-SVM, HOG-KNN, 3D SIFT-SVM and 3D SIFT-KNN. Results are displayed in following tables.

Descriptor/ Classifier	3D SIFT				
	k/Data	100	500	800	1000
SVM	Training	95.83	95.42	95.42	95.63
	Testing	43.33	38.33	46.67	61.67
KNN	Training	95.21	94.79	94.17	94.37
	Testing	32.50	36.67	42.50	49.17

Table I: HOG action recognition accuracy for training and testing videos using SVM and KNN classifiers

Descriptor/ Classifier	HOG				
	k/Data	100	500	800	1000
SVM	Training	96.04	98.75	98.54	**99.17**
	Testing	82.5	**85.83**	85.00	85.00
KNN	Training	98.54	98.96	97.71	97.50
	Testing	80.00	75.83	66.67	69.17

Table II: 3D SIFT action recognition accuracy for training and testing videos using SVM and KNN classifiers

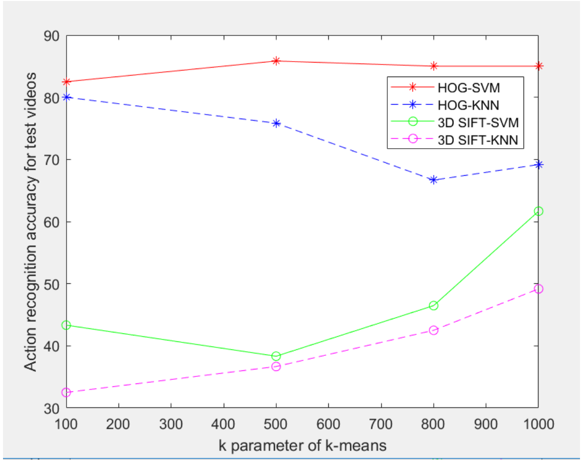

Fig. 3: Graph of Accuracy vs k-parameter of k-means clustering algorithm for test videos

It can be observed from the graph that with increase in k value the accuracy of HOG-SVM combination increases then decreases slightly and becomes steady while that of HOG-KNN decreases first then increases slightly. Accuracy of 3D SIFT-SVM combination shows first decrease and then steep rise while that of 3D SIFT-KNN shows a continuous rise. Among the four combinations HOG-SVM clearly shows superior performance. The maximum testing accuracy is obtained at 500 k parameter value of k-means for HOG-SVM combination among the four k parameter values used for experiments. The maximum training accuracy is obtained at 1000 k parameter value of k-means for HOG-SVM among the four k parameter values used for experiments.

V. Conclusion and Future Work

In this paper, experimental evaluation of four combinations of two feature descriptors and two classifiers is done. The objective is to find the superior combination for future research in video-based action recognition problem. The best performance on KTH dataset has been achieved with the HOG-SVM combination among the four combinations. It is observed that the k parameter of the k-means clustering algorithm is an important parameter which has an impact on the classification performance. Increasing the size of feature vector of the 3D SIFT may improve the performance. Due to usage of videos the computation time required was very high. It can be further reduced by applying parallel computations.

References

1. N. Nayak, R. Sethi, B. Song, and A. Roy-Chowdhury. Motion pattern analysis for modeling and recognition of complex human activities. In Guide to Video Analysis of Humans: Looking at People. Springer, 2011

2. Zhu, Guangyu, Ming Yang, Kai Yu, Wei Xu, and Yihong Gong. "Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor", Proceedings of the seventeen ACM international conference on Multimedia - MM 09 MM 09, 2009

3. Lu, Nan, Jihong Wang, and Q.H. Wu. "An optical flow and inter-frame block-based histogram correlation method for moving object detection", International Journal of Modelling Identification and Control, 2010.

4. D.G. Lowe, “Distinctive Image Features from Scale Invariant Keypoints”, in International Journal of Computer Vision (IJCV), 2004

5. 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.

6. C. Schüldt, I. Laptev, and B. Caputo. "Recognizing human actions: a local SVM approach", Proceedings of the 17th International Conference on Pattern Recognition 2004 ICPR 2004, 2004.

7. Navneet Dalal, Bill Triggs, “Histograms of oriented gradients for human detection”, Computer Vision and Pattern Recognition, 2005.

8. Alexander Klaser, Marcin Marszalek, Cordelia Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. Mark Everingham and Chris Needham and Roberto Fraile. BMVC 2008 19th British Machine Vision Conference, Sep 2008, Leeds, United Kingdom. British Machine Vision Association, pp.275:1-10, 2008. <inria-00514853>

9. 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.

10. Christopher Bishop, “Pattern Recognition and Machine Learning”, Springer, 2007

11. G.Akilandasowmya, P.Sathiya, P AnandhaKumar, “Human Action Analysis using KNN Classifier", Seventh International Conference on Advanced Computing (ICoAC),2015.

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

13. Jyoti Yadav,Monika Sharma,"A Review of K-mean Algorithm", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- 2013