Semi-Supervised Human Action Recognition Using Clustering and SVM Classifier

G.Gowri Pushpa [1]***, V.Parimala*** [2]***, V.Divya*** [3]***, K.Leela Prasad*** [4]***, G.Gopalakrishna*** [5]

[1]*Assistant Professor, Dept of Computer Science and Engineering, Anil Neerukonda Institutes of Technology and Sciences, Visakhapatnam, India*

[2345]*Students, Dept of Computer Science and Engineering, Anil Neerukonda Institutes of Technology and Sciences, Visakhapatnam , India*

Abstract : In recent years, Human action recognition gained more importance for its variety of applications like video surveillance, Entertainment, Human-Robot Interaction and so on. Action recognition targets recognising different actions from a sequence of observations and different environmental conditions. This paper compares two different handcrafted feature extraction techniques (i.e..SIFT and optical flow) with SVM classifier on KTH dataset. The SIFT technique detects the interest points for each frame in the video. Then building a codebook using Bag-of-Words approach with the clustering technique as K-means clustering, then the classification is done using the SVM classifier. Similar, process is followed for another feature extraction technique i.e. optical flow.The experimental results of our approach gives an accuracy of 80.75% for SIFT and 83.45% for optical flow.

Key words : Action representation, Feature extraction, SIFT, Optical Flow, Key Points, Descriptors, K-means clustering, Bag-of-Words, Action classification, SVM classifier.

1 INTRODUCTION

Vision-based Human action recognition has received increasing attention in computer vision and has made significant progress in recent years. Adequate training data is essential for detecting specific human actions. This type of recognition not only provides us with knowledge about certain people, but it also assists us in understanding their intentions, feelings, and attitudes. In this paper, Human actions are recognised by extracting features from videos using the built-in SIFT[5] and Optical Flow[6] feature extraction algorithms. Our proposed system is trained and evaluated on KTH dataset[1] followed by extracting the features using the hand-crafted feature extraction techniques, then the output features are divided into training and testing sets, the training set is used for forming the clusters using K-Means clustering algorithm with different values of ‘k’.These clusters form a vocabulary using bag-of words[2,4] vector and the linear SVM classifier is trained with these vectors. Then the testing set is used for evaluation which at last gives the confusion matrix which is used for calculating the accuracy of the model.

2 RELATED WORK

Human Action Recognition approach mainly focuses on video representations, by extracting the features and describing their formats, there are different feature extraction techniques are available so there is a need to choose the efficient and relevant one. Vision based human action recognition involves feature extraction and action classification by assigning labels to the classes. The assignment of labels is a challenging task as the classifier can misclassify the actions because of the challenges like inter and intra- class variations, Cluttered Background and Camera Motion and Insufficient annotated data etc[2]. In [8], tracking of keypoints generated from SIFT flow trajectory algorithm and descriptor are trained for various trajectories. There are three different types of classifiers: The method of the Bag-Of-Words, Linear and Non-Linear SVM, results of these methods were compared and tabulated. These different techniques were implemented on UCF Sport Action dataset. Two-third video clips of each ten different actions were taken as training dataset and remaining were used for testing purpose. Linear SVM classification showed 87% of accuracy whereas in Non-linear SVM this accuracy was reduced to approximately 80%. Thus, in this proposed method Linear SVM classifier is used for better performance.

3 DATASET

The dataset used in this model is KTH dataset[1] which contains videos of 6 action categories i.e., boxing, hand-clapping, hand-waving, jogging, running, walking performed in 4 different scenarios i.e., outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3, indoors s4 as shown in Fig 1. For each scenario, there are 25 videos, each category 100 videos, thus, the dataset contains 600 videos. All videos have a frame rate of 25fps and a spatial resolution of 160x120 pixels.All the videos are stored in the AVI format. This dataset after the feature extraction is divided into training and testing sets, where the classifiers are trained on the training set and the results are obtained on the test set.

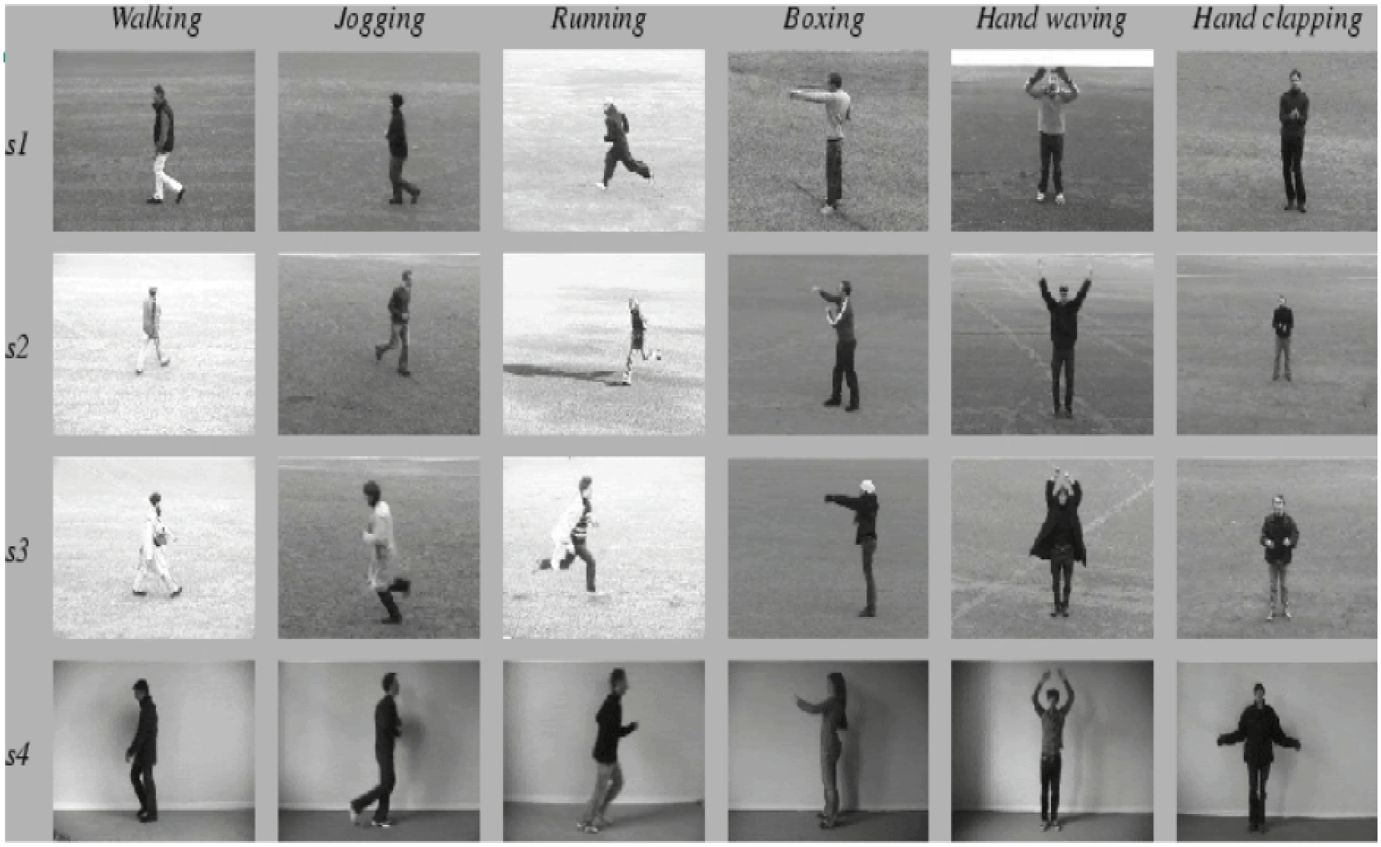

Fig-1: Action database: different actions and scenarios

3 PROPOSED METHOD

Human Action Recognition involves two steps, Action representation and Action classification[2]. The main aim of Action representation method is to convert the input action video into a feature vector and extract representative and discriminative information from the human actions. The input videos may differ in their motion speed, camera angle and pose variations, the selection of action representation method is to maximaize the discrepancy between the actions and the Action classification method concludes the action label from the vector. In this proposed system, Bag-of-words approach is used to classify the actions, in which there is a use of clustering algorithms and quantization techniques to minimize the noise and classification error.

3.1 Approach

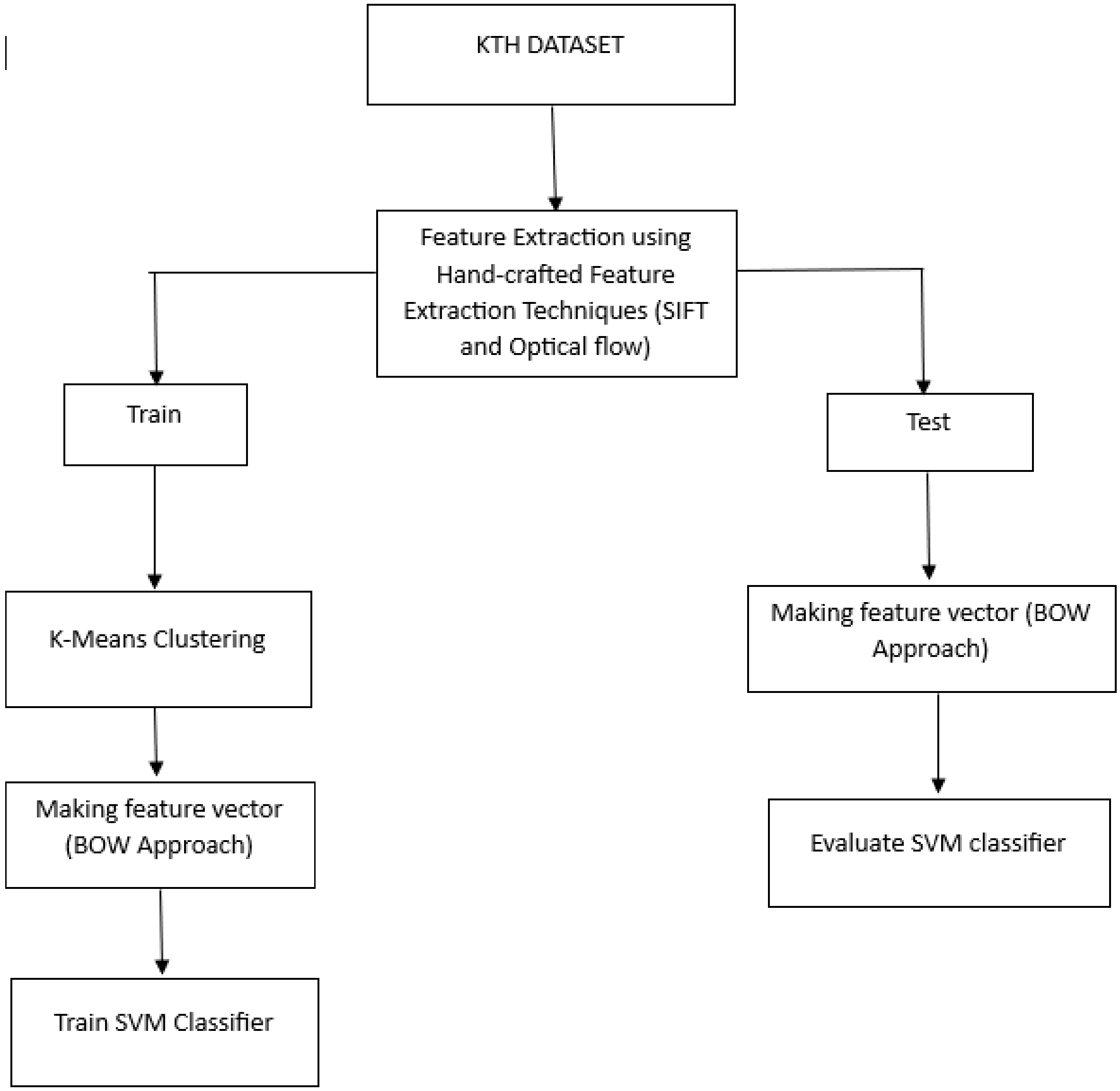
KTH DATASET
Feature Extraction using Hand-crafted Feature Extraction Techniques (SIFT and Optical flow)
Train
Test
K-Means Clustering
Making feature vector (BOW Approach)
Making feature vector (BOW Approach)
Evaluate SVM classifier
Train SVM Classifier

Fig-2: Proposed System Approach

3.1 Action Representation

The attributes like motion speed, camera angle, look and pose variations of human action may vary. This makes representation of action difficult. The main aim of action representation is to convert an action video into a feature vector , thereby minimizing the variations and extracting meaningful information. Here, we focus on hand-crafted action representation methods like SIFT and Optical Flow.

3.1.1 Feature Extraction Techniques

a) SIFT Feature Extraction Technique

SIFT stands for Scale-Invariant Feature Transform[5] which detects the features that are scale and rotation invariant. This method is used to extract features from the dataset. This method comprises of 3 steps :

(i) Scale space extrema detection : The goal of this stage is to point the position of keypoints on the image by finding extreme values in scale space. The pixel will be compared to 26 pixels, 8 pixels in the same scale and 92 pixels in the scale before and after.

(ii) Keypoint localization and orientation assignment : The direction and magnitude of each keypoint will be determined by the gradient direction of its nearby pixels. Then, as the keypoint region, identify and locate the most prominent orientation in the region. The effect of orientation is essentially canceled, making rotation invariant.

(iii) Keypoint descriptor : The SIFT feature vector is created at this point. In the actual calculation, describe each keypoint using 4x4 matrix to improve matching stability. For each keypoint, 128 data points, or a 128-dimensional SIFT vector, are retrieved. SIFT vector is no longer affected by geometric adjustments such as scaling and rotation.

b) Optical Flow

Optical flow technique[6] describes the image motion. The video is divided into frames and when optical flow technique is applied on those frames, it calculates the velocity for the points in the frames. It compares the previous frame and the next frame and then computes the intensity difference among the points in the frame to pair. Once the pairing is done, it calculates the velocity by computing the distance that the point has moved.

3.1.2 Bag Of Words Model

After all the keypoints are extracted from the dataset, now build the vocabulary using those keypoints to train the SVM classifier. Bag of words model is generally used to represent textual information where it counts the number of repeating words in a text. Similarly, when it comes to keypoints, all the feature points are considered as a visual words and the K-Means algorithm is used to construct these visual words. Using these words, we need to construct vocabulary of the BOW model.

This step helps to reduce the computation and increase the classification accuracy. Here, we are using the K-Means algorithm for different values ‘k’ such that the accuracy may differ for different numbers of clusters. When it comes to constructing vocabulary we used vector quantization(vq()) and Tf-Idf weighting scheme.Vector quantization (vq()) method compares each observation with the cluster centroids and assigns the observation to the closest cluster. Tf-Idf stands for Term Frequency and Inverse Document Frequency, where the term frequency is used to calculate the frequency of the word and Inverse Document frequency is to find the number of occurrences of the term in the clusters, which reduces computation complexity.

3.2 Action Classification using SVM classifier

Support Vector Machine (SVM) is a popular classification method which is used now-a-days for various classification problems. It is a supervised machine learning algorithm of 2 types i.e., linear and nonlinear, here in this system we used linear SVM classifier. SVM classifier works well for high dimensionality data. The main idea of SVM is to find the nearest points to the hyperplane and find the optimal solution using its parameters.

4 EXPERIMENTAL ANALYSIS AND RESULTS

From the KTH dataset, out of 600 videos, 384 videos are considered for training set and remaining 216 videos for testing as per [1]. The equation for measuring the accuracy is:

Accuracy = number of testing videos classified correct/total number of testing videos * 100%

For SIFT and optical flow, we have experimented with different numbers of clusters.The major purpose of this section is to determine how many clusters are suited well enough to increase the accuracy of the grayscale image using SIFT. The following are the results:

(i) Using SIFT Technique

S.No	‘K’ Value	Overall Accuracy
1	10	27.53%
2	100	50.26%
3	200	59.68%
4	400	72.25%
5	800	80.75%

Table-1: Results obtained using SIFT

(ii) Using Optical Flow Technique

S.No	‘K’ value	Overall Accuracy
1	10	31.32%
2	100	54.87%
3	200	78.16%
4	400	81.99%
5	800	83.45%

Table-2 : Results obtained using Optical Flow

5 CONCLUSION AND FUTURE SCOPE

Human Action Recognition proposed in this paper provides an effective solution to get an idea of which feature extraction method gives the better accuracy. From the above analysis we can say that using optical flow feature extraction technique gives better results than SIFT feature extraction technique.

This scope of study will expand in future to recognize more human actions which can be used for video surveillance, to predict any malicious activities are being done.

6 REFERENCES

[1] Schuldt, C.; Laptev, I.; Caputo, B. (2004). [IEEE Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. - Cambridge, UK (2004.08.26-2004.08.26)] Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. - Recognizing human actions: a local SVM approach. , (), 32–36 Vol.3. doi:10.1109/ICPR.2004.1334462

[2] Yu Kong, Member, IEEE, and Yun Fu, Senior Member,IEEE, " Human Action Recognition and Prediction : A Survey" Journal of Latex class files, Vol-13,No - 9 , September 2018

[3] Qazi, Hassaan Ali; Jahangir, Umar; Yousuf, Bilal M; Noor, Aqib (2017). [IEEE 2017 International Conference on Information and Communication Technologies (ICICT) - Karachi, Pakistan (2017.12.30-2017.12.31)] 2017 International Conference on Information and Communication Technologies (ICICT) - Human action recognition using SIFT and HOG method. , (), 6–10. doi:10.1109/ICICT.2017.8320156

[4] Li, Qilong; Wang, Xiaohong (2018). [IEEE 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) - Singapore (2018.6.6-2018.6.8)] 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) - Image Classification Based on SIFT and SVM. , (), 762–765. doi:10.1109/ICIS.2018.8466432

[5] Zhang, Jia-Tao; Tsoi, Ah-Chung; Lo, Sio-Long (2014). [IEEE 2014 International Joint Conference on Neural Networks (IJCNN) - Beijing, China (2014.7.6-2014.7.11)] 2014 International Joint Conference on Neural Networks (IJCNN) - Scale Invariant Feature Transform Flow trajectory approach with applications to human action recognition. , (), 1197–1204. doi:10.1109/ijcnn.2014.6889596

[6] Danafar, S., & Gheissari, N. (n.d.). Action Recognition for Surveillance Applications Using Optic Flow and SVM. Lecture Notes in Computer Science, 457–466. doi:10.1007/978-3-540-76390-1_45

[7] Nagabhushan, T.N.; Aradhya, V. N. Manjunath; Jagadeesh, Prabhudev; Shukla, Seema; M.L., Chayadevi (2018). [Communications in Computer and Information Science] Cognitive Computing and Information Processing Volume 801 || Human Action Detection and Recognition Using SIFT and SVM. , 10.1007/978-981-10-9059-2(Chapter 42), 475–491. doi:10.1007/978-981-10-9059-2_42

[8] Zhang, Jia-Tao; Tsoi, Ah-Chung; Lo, Sio-Long (2014). [IEEE 2014 International Joint Conference on Neural Networks (IJCNN) - Beijing, China (2014.7.6-2014.7.11)] 2014 International Joint Conference on Neural Networks (IJCNN) - Scale Invariant Feature Transform Flow trajectory approach with applications to human action recognition. , (), 1197–1204. doi:10.1109/ijcnn.2014.6889596

[9] Kumar, S. Santhosh; John, Mala (2016). [IEEE 2016 International Carnahan Conference on Security Technology (ICCST) - Orlando, FL, USA (2016.10.24-2016.10.27)] 2016 IEEE International Carnahan Conference on Security Technology (ICCST) - Human activity recognition using optical flow based feature set. , (), 1–5. doi:10.1109/CCST.2016.7815694

[10] Manosha Chathuramali, K. G.; Rodrigo, Ranga (2012). [IEEE 2012 International Conference on Advances in ICT for Emerging Regions (ICTer) - Colombo, Western, Sri Lanka (2012.12.12-2012.12.15)] International Conference on Advances in ICT for Emerging Regions (ICTer2012) - Faster human activity recognition with SVM. , (), 197–203. doi:10.1109/ICTer.2012.6421415

[11] Das Dawn, Debapratim; Shaikh, Soharab Hossain (2015). A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. The Visual Computer, (), –. doi:10.1007/s00371-015-1066-2