International Journal of Applied Engineering Research ISSN 0973-4562 Volume 15, Number 5 (2020) pp. 460-467
© Research India Publications. http://www.ripublication.com

Human Action Recognition using 2 Level Classification Method

Ravina Kumari,
M.tech Schoalr (ECE),
N.C College of Engineering and Technology,
India.

Jagtar Singh,
Assistant Professor (ECE),
N.C College of Engineering and Technology,
India.

Abstract

Human Activity Recognition (HAR) is a popular topic for researchers. The many researchers presented the HAR for image classifications. The HAR from the video samples is suffered from various challenges like cluttered background, brightness and motion variation. In this dissertation, we proposed a Two-level classification model for human activity recognition. We take the UT interaction video dataset to validate the performance of the work. The UT interaction dataset consists of 20 sequence videos with the six different interaction category of the human-human interaction. The input video segmented into the frames using a python script. The global features are evaluated with the Histogram of Gradient (HOG) and Histogram of Optical Flow (HOF). A features vector obtained after the feature extraction from the frames of the video. The dimension of the HOF and HOG feature vector is reduced by the Principle Component Analysis (PCA) method. The extracted features are fed to the SVM classifier which provided two different classes of six human-human interaction action. Class 1 contains the fighting interaction like Kicking, Punching and Pushing and Class 2 contains the normal interaction like Pointing, Hugging and Hand Shaking.

Again features are extracted from the two classes of human interaction. The Speed up Robust Features (SURF) is used to detect the motion features from the classified categories of interaction. The detected features are represented into the visual words by using the Bag of Visual Word (BoVW) approach. The BoVW approach encoded the features or key points extracted from the segmented frame. The output of SURF-BoVW features is fed to the Neural Network classifier, which classifies the human-human interaction. The proposed method provided efficient results than the single level classier like SVM and KNN. The proposed method provided approximate 99 % classification accuracy for the UT interaction dataset.

Keywords- SURF, BoVW, HOG, HOF, SVM and Neural Network

I. INTRODUCTION

In modern days, Human Activity Recognition (HAR) is a popular topic for researchers. The visual-based HAR has wide applications in the computer field as well as in the machine learning field. The function of HAR is that understand the human action from the videos or images automatically. Action recognition is a difficult task due face some challenges in the HAR. The problems include variation in-camera, blur background, human body variation, stationary and moving object, motion and illumination level at different stages affects the HAR phenomena. The intensity of these problems may vary with respect to the activity which is to be tested. Based on the general perspective, human activity is divided into four different categories. The categories are actions, interactions, group activities and gestures.

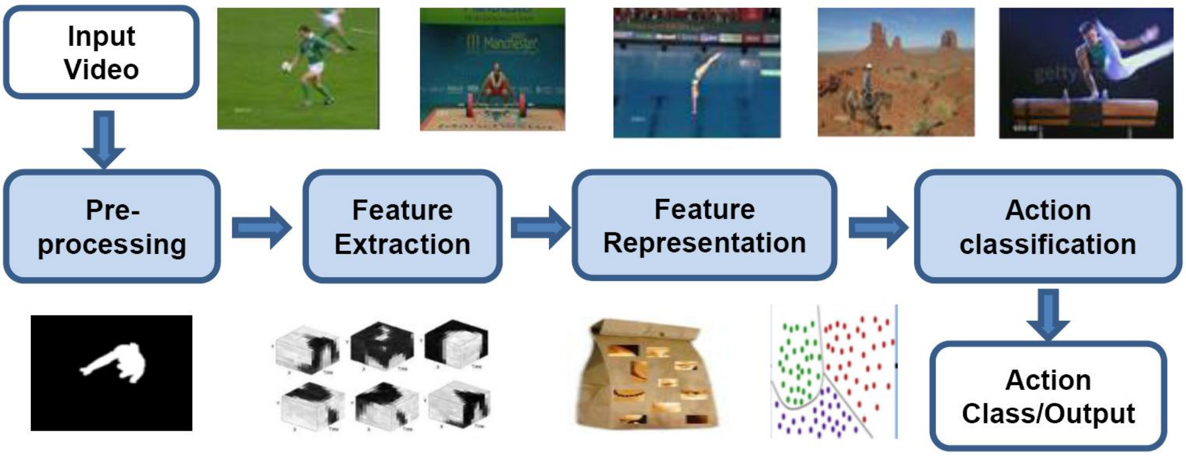


Figure 1. Schematic diagram of HAR

Figure 1 shows the schematic arrangement of a human activity recognition system. The input can be a video or image in which the person performs some activity. The first input is preprocessed means noise or interruption is removed in preprocessing. Next step isss featttureees extacrtion in which phase motion features or temporal features are extracted from the image or videos. A feature vector is designed to represent the extracted features in the frequency domain. Classifiers received the input from the feature vector and classified the human action. The output is the recognized activity of the human being.

1.1 Overview of Research

In this work, we performed human activity recognition using the UT interaction dataset. The UT interaction dataset consists of videos of human-human interaction in six different classes. Total of 20 video sequences present in the dataset. The six different categories of human interactions are punching, kicking, pushing, shake hands, point and hugging. The HOG and HOF (Histogram oriented Optical flow) motion descriptor is used to identify the motion features. Both cases are used to detect the motion form the UT interaction database videos. Find the global features vector by combining HOG [1,3] and HOF. The length of the global features vector is not in a symmetrical manner, so the Principle Component Analysis (PCA) [3] scheme is used to reduce the dimension of the global feature vector. Generate the feature descriptor by concatenating PCA modified HOG and HOF. The two-level classifier performs the action classification. SVM is used as the first classifier, which classifies the frames of the video into two categories.

The two classes are class A and class B, which further converted into the frame section. The features of the two classes are extracted by the Speed up Robust Features (SURF) and Bag of Visual Words (BoVW) method. Both methods of feature extraction have high speed and accurate performance than the existing approach. The detailed about the SURF and BoVW will be provided in section 3. The features representation is performed by the DWT (Discrete Wavelet transform). A neural network classifier is used to classify the activity of human interaction. The SURF and BoVW features trained the NN classifier and provided the action classification of the human interaction dataset.

1.2 Objectives

In this work, we classified the UT interaction video dataset using a two-level classification method. The motion features extracted by the HOG and HOF approach, as we studied in [2]. The SURF –BOW based approach is used to identify the features form the frames of the video after the motion features detection. Then two-level classifier is used to classify the different categories of human action. The following objectives are;

1. To extract the features form the UT interaction database videos using HOG and HOF motion feature descriptor and SURF-BWO for action feature vector extraction.
2. To classify the human action by using two-level classifier SVM and Neural Network
3. To compare the results with a single level classifier approach.

II. RELATED WORK

We studied lots of work related to human activity recognition. The features based approaches are most popular for the HAR task. The SURF and BoVW combination is proposed in many studies like [1, 2, 13, 14, and so on]. The SIFT approach provided a better accuracy of action recognition. The Support Vector Machine (SVM) is used as a classifier in most of the researches. Some Spatio-temporal based features recognition approach like HOG, HOF, MBH, and velocity are provided in different studies. The proposed methods are used to test the large datasets as well as the small datasets and real-world video monitoring. The KTH [1] and UCF [5] sports data are analyzed in most studies.

After studied various approaches of HAR, we get an idea of using the SURF and BoVW approach for the features recognition of human activity. We can be tested the UT interaction datasets with the two-level classification approach. The SVM and NN are used as the classifier, which classifies the different human-human interaction actions.

III. DATASET AND FEATURES EXTRACTOR

There are various human activity and interaction datasets available publically. Some publically available datasets can be used for the comparison among the different approaches. The KTH dataset [1] used most popularly for human activity recognition. KTH consists of 6 classes of 408 recorded videos.

3.1 UT interaction dataset

The UT interaction dataset consist of 20 sequence videos of six classes of human-human interactions. The length of each video is near about 1 minute. The interactions are handshaking, pushing, punching, pointing, kicking and hugging. In each video, at least one human-human interaction execution is provided in the UT interaction dataset. There are six different human-human interactions execution are obtained per video on average. In the videos, various persons participate with their 15 different clothing conditions shown in the sequence manner. The resolution of the videos is 720×480, and 200 pixels values show the height of a human in the video. The labels are provided in the ground truth manner with bounding boxes and time intervals [31].

The 20 videos are divided into two sets, Set 1 and Set2. Set 1 contains the ten videos which are captured in the parking lot. The set 1 sequence videos are having different zoom rate and static background condition using small camera jitter [31].

3.2 Speed up Robust Features (SURF)

The speed-up robust features (SURF) is a local feature detector and descriptor method. This method is commonly used in object recognition and classification applications. The SIFT descriptor inspires the SURF working principle. SURF provides better performance than the SIFT approach. The main applications of the SURF process are object detection, face detection, and extract interest points from the motion of the video. The modification in the SIFT version is known as the SURF. The execution speed is also improved by using the SURF [32].

The location of object and recognition is computed using the SURF descriptor. It tracks the objects very efficiently and extracts the point of interest. Herbert Bay develops the concept of SURF in 2006 for image features classification [32].

3.2.1 SURF descriptor

The SURF descriptor provides a robust description of image features. The description features representation should be unique. The description of each interest point is identified locally. The range or dimension of the descriptor has directly affected the robustness and computational complexity. The short-range of the descriptor is more accurate and efficient with appearance variations.

The Haar wavelet response in both directions x and y-axis within the circular radius around the point of interest is computed, and the point at which point of interest is detected is known as scale. The output response is weighted by the Gaussian function centered at the point of interest [32]. A square region alignment is constructed on the selected region and extracts the SURF descriptor form them.

3.3 Bag of Visual word (BoVW)

The extracted features from the image are treated like the words in the bag of a visual-based approach. It is the count of words sparse vector over the vocabulary. It is a common method of image classification. In the BoVW approach, the total number of words is count, which shown in the image pattern and uses their frequency to understand the word in the document with histogram representation. The image features are used as the word of the document and available in the unique form [1].

The three main steps feature detection, features description and codebook generation are to be followed for the implementation of the BoVW approach. The feature detection is that the task in which image information and key points are identified. The outcomes are in the form of image features. After the features detection, the features representation task takes place. During features representation, each image or frame subtract from the local patches. The patches are represented by the numerical vectors, which are known as the features descriptors.

IV. PROPOSED METHOD

We proposed a Two-Level classification approach for the action recognition of human-human interaction using the UT interaction dataset. The video sample is converted into the frames in the preprocessing of the dataset. The frames of the video consist of several global features, which are figure the Histogram of Gradient (HOG) and Histogram of Optical Flow (HOF) approaches. The HOG and HOF features sets have a larger dimension, which is reduced by the Principle of Component Analysis (PCA) method. The low dimension features set HOG and HOF are concatenate for the production of feature descriptors. The feature descriptor detects the features of interest points. The Support Vector Machine (SVM) classifier is trained with the HOG and HOF features descriptor outcomes. SVM act as the first classifier, which classifies the sequence videos of six activities into two sections.

The first section is represented by Set A, which has the first three (1, 2 and 3) activity sampled video frames. The section is Set B, which contains the last three (4, 5 and 6) activities of the human-human interaction. The Bag of the visual word (BoVW) and Speed Up robust features (SURF) based approach is applied to each classified frame of the video. The SURF is used for extraction purposes from the segmented sequence videos. The extracted features are represented in the features set vector. The descriptor analyzed the features vector and provided input to the BoVW. The bag of the visual word provided the coding to the extracted features and minimized the distance of the interest points. The K-Mean clustering algorithm is used in the BoVW to reduce the distance among the interest points of scatter. The extracted feature sets are used to train the second classifier Neural Network (NN).

4.1 Steps of proposed work

The entire work is provided into the following steps;

1. Convert video into the frame- We developed python code for converting the sequence videos into the frames. The UT interaction dataset contains the sequence of video data. The classification task is not performed with the motion data. So frames are extracted from the sequence videos using a python script.
2. The frames are represented into the Histogram manner by extracting the HOG, and HOF features vectors: the key points and interest points extracted from frames in this step. The global features are computed using the HOG and HOF approach.
3. The length of the HOG and HOF features vector is minimized by the Principle Component Analysis method.
4. Then Concatenate HOG and HOF and produced a features descriptor. The features descriptor extracts the features like texture and visual from the frames.
5. By using the features, descriptors train the Support Vector Machine (SVM) classifier for the classification of similar human interaction categories among six interactions action.
6. The data frames dived into the two classes; Class A and Class B. Class A contain the interaction category kicking, pushing and punching, and Class B consists of the action interaction category pointing, handshaking and hugging. It is called level one classification.
7. Evaluate the SURF features from the frames of the classified video. The SURF descriptor computes the action and texture features. The visual words are designed for the SURF features using the Bag of visual word approach
8. The edges and corner features are computed in these steps using a Discrete Wavelet Transform.
9. The SURF-BoVW extracted features are used to train the Neural Network. The NN classifier recognizes the human interaction action from the given dataset.
10. Compare the results of the proposed method (Two Level classifier) with the SVM and KNN classifier. The proposed method obtained higher classification accuracy than the SVM and KNN

4.2 Flowchart of proposed work

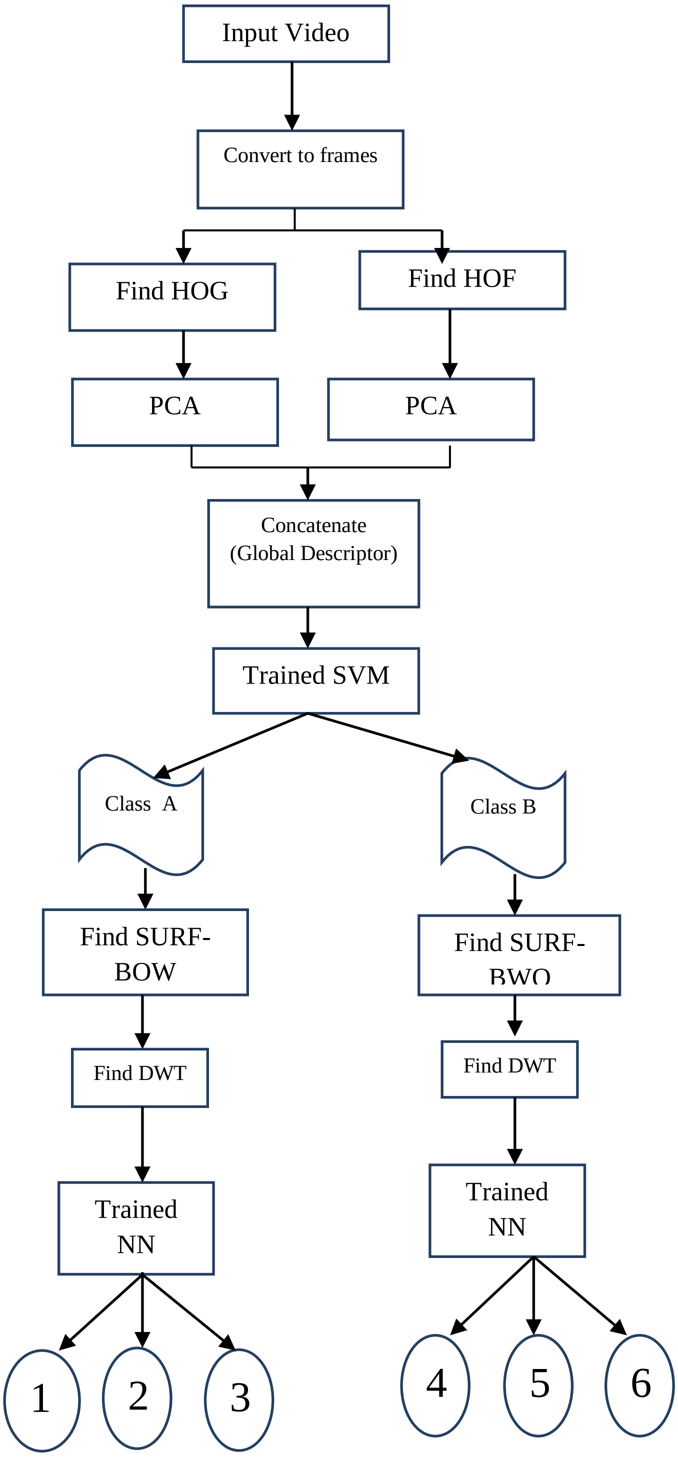


Figure 2. Flowchart of proposed work

V. RESULTS AND DISCUSSION

We tested the proposed method (2-level classification) using (SVM-NN) on the UT interaction dataset. The UT interaction contains the video sequence data which converted into the frames using the python code. The collected frames are defined the interaction activity of the human. The key points and interest points represented into the features vector using the HOG and HOF features. The features representation is presented in the histogram manner. These histogram features are fed to the 1st level classifier SVM. It segmented the frames as per their related activity. Two categories are classified with the 1st level classifier. The categories are fighting and friendly behavior based. In the fighting category the kicking, punching and pushing interaction of human arrived and other category contains the pointing, hand shaking, and hugging interaction.

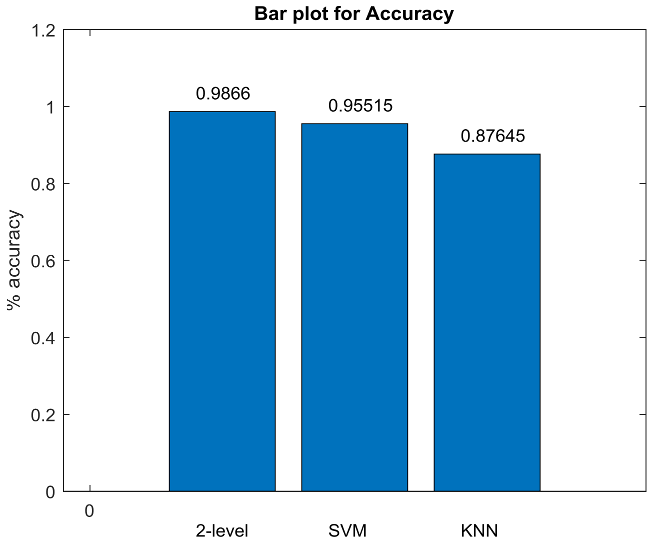


Figure 3. Accuracy comparison among the 2-level, SVM and KNN classifiers

The performance evaluation of the proposed method is done with four configurations like accuracy, precision, recall and confusion matrix. The classification accuracy should be maximum for the proposed work. The higher classification accuracy reflects the betterment of the work. Figure 3 shows the accuracy comparison of the proposed method, and single classifiers (SVM and KNN). The human interaction action is more accurately classified by the 2 level classification than the SVM and KNN. The accuracy of the proposed method is nearly to the 99%. The single level classifier is provided accuracy less than the proposed method. The SVM classifier provided the 95.55 classification accuracy and KNN classifier provided 88%.

The classification accuracy is affected by the number features extracted by the HOG, HOF and SURF features. More frames are correct identified in the proposed method evaluations. In single level the features extraction and classification task is performed once so they do not provided better accuracy.

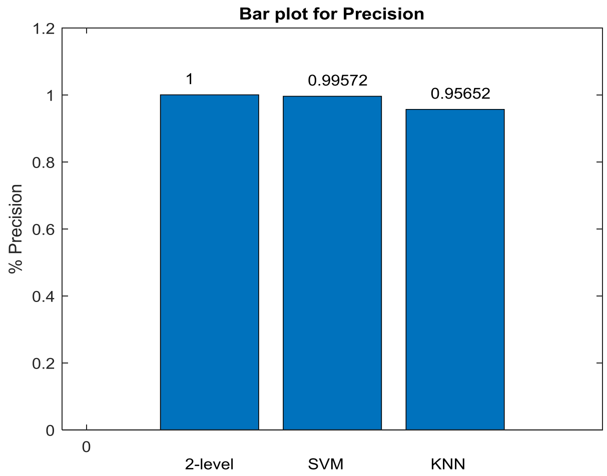


Figure 4. Precision comparison among the 2-Level, SVM and KNN classifiers

Figure 4 provided the comparison among the precision plot of 2-level classification, SVM and KNN classifiers. The precision is the performance evaluation parameters of classification action in HAR. The value of precision is placed between the 0 to 1. The maximum value of precision is achieved in the 2-level classification method. The proposed method provided the higher precision than the SVM and KNN classifiers. Table 1 shows the comparison among various method of classification.

Table 1. Comparisons observation of 2-level, SVM and KNN approach

Evaluation parameters	Accuracy	Precision	Recall
2-Level classification	98.66	1	0.9945
SVM	0.9551	0.9957	0.9660
KNN	0.8764	0.9565	0.9543

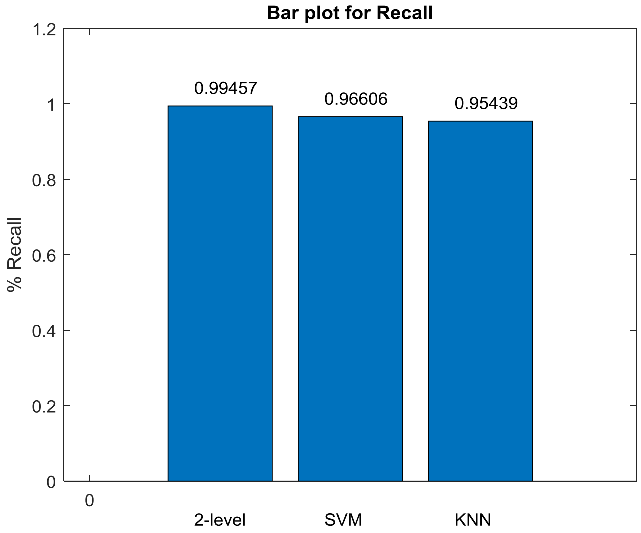
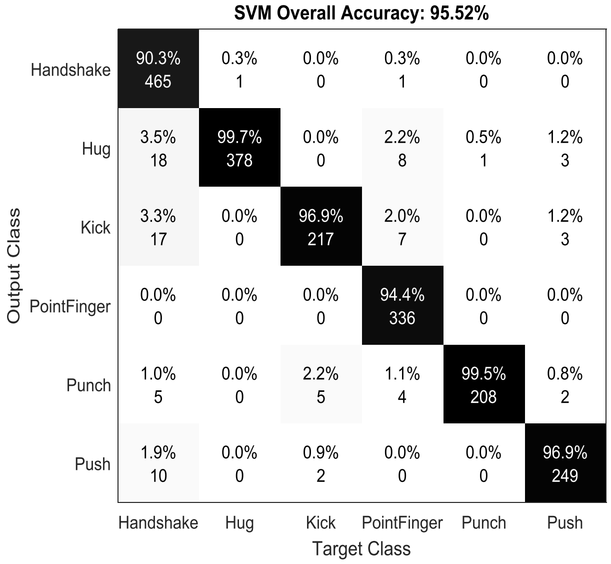


Figure 5 Recall comparison among the 2-level, SVM and KNN classifiers



	Handshake	Hug	Kick	PointFinger	Punch	Push
Handshake	90.3% 465	0.3% 1	0.0% 0	0.3% 1	0.0% 0	0.0% 0
Hug	3.5% 18	99.7% 378	0.0% 0	2.2% 8	0.5% 1	1.2% 3
Kick	3.3% 17	0.0% 0	96.9% 217	2.0% 7	0.0% 0	1.2% 3
PointFinger	0.0% 0	0.0% 0	0.0% 0	94.4% 336	0.0% 0	0.0% 0
Punch	1.0% 5	0.0% 0	2.2% 5	1.1% 4	99.5% 208	0.8% 2
Push	1.9% 10	0.0% 0	0.9% 2	0.0% 0	0.0% 0	96.9% 249

Figure 6 Confusion matrix plot for the SVM classification HAR

Figure 5 shows the recall parameters comparison among the 2-level, SVM and KNN classifiers. The higher value of recall is good for the classification purpose of human interaction. In the interaction classification the proposed method recall is better than the SVM and KNN classifiers. The recall parameters are directly related to the classification accuracy of the system. The greater value of recall proves that higher accuracy achieved. In this comparison the 2-level classification model provided 0.9964 value of the recall parameters which is greater than the

The confusion matrix plot is also used to measure the accuracy of interaction wise case. This plot provided the accuracy with the samples corrected or not. The classification accuracy for the each interaction action activities provided better analysis for the samples. Figure 6 shows the matrix plot for the single level classifier SVM. Only two interaction action Punch and Hug provided better accuracy of classification. The misclassification rate is higher than the 2-level classification model. The less accuracy means the low features are extracted for that action interaction. The SVM confusion matrix plot shows the average accuracy is 95.55% which is much lesser than the proposed method.

Figure 7 shows the confusion matrix plot for the proposed method. The number of samples of the human interaction plays an important role on to the classification accuracy. In this case point figure and push interaction activities provided 100% classification accuracy. The main reason behind this is using the dual classifier SVM and NN for the classification. If the number of samples are more in an any action then the improved accuracy is achieved.

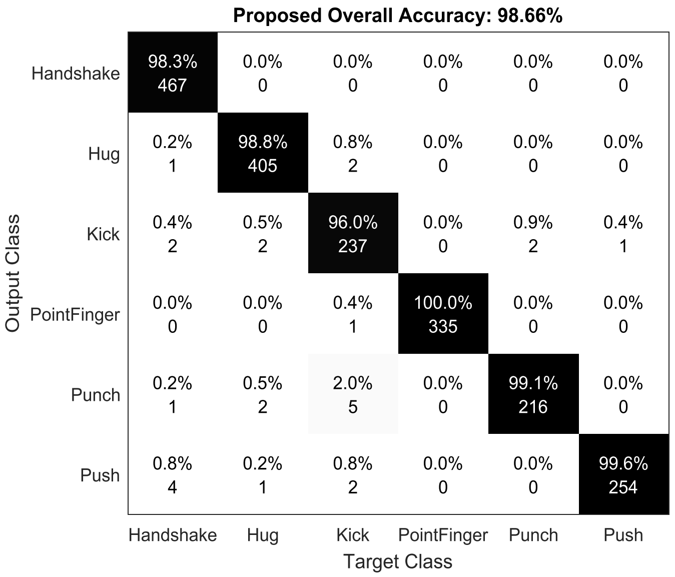


Figure 7 Confusion matrix for proposed method accuracy

Table 2. Performance accuracy as per the interaction action

Interaction activities	2-Level classification (%)	SVM (%)	KNN (%)
Hand shake	98.3	90.3	88.4
Hug	98.8	99.7	84.9
Kick	96	96.9	84.6
Point Finger	100	94.4	88.8
Punch	99.1	99.5	93.2
Push	99.6	96.9	87.7
Overall	98.66	95.52	87.64

Figure 8 shows the confusion matrix plot for the KNN classifier method. The average classification accuracy is 87.64 in case of KNN classifier based approach.

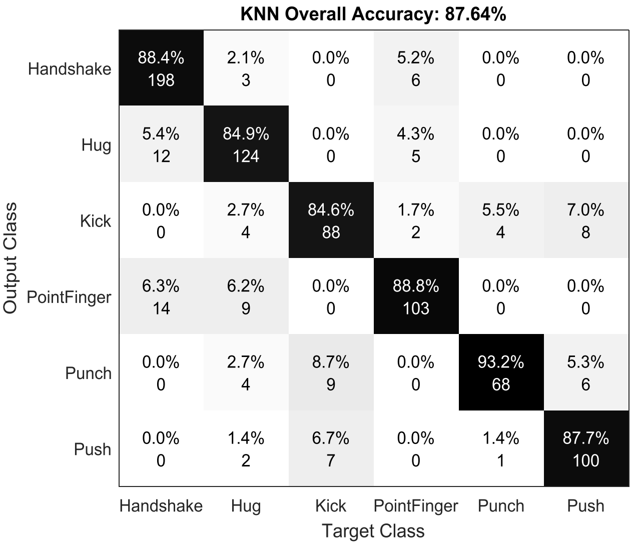


Figure 8 Confusion matrix for the KNN accuracy

So the proposed method provided better accuracy then the single level classifiers SVM and KNN.

VI. CONCLUSION

In this work, we proposed a Two-Level Classifier for human activity recognition. The input is considered as the sequence of video samples. We use the UT interaction dataset which contains 20 sequence videos with the six different human interaction activities. The video samples are segmented into the frames than Histogram features HOG, and HOF is computed from them. The global features are obtained from the HOG and HOF, and the PCA method is used to reduce the dimension of the features vector. The output of features vector fed to the SVM classifier which provided two classes of human interaction. One class contains the fighting behavior interaction action and the second class contains the natural behavior interaction action of a human. We used neural network classifier as the second level classification. For the second level classifier, we used the SURF-BoVW approach to extract the features from the classified interaction activity. The encoded features are provided to the NN which identified the interaction activity from the signal. The two-level human activity classification approach provided better performance than the single level classifiers like SVM and KNN. The proposed method provided approximate 99 % classification accuracy for the UT interaction dataset.

REFERENCES

[1] Aslan, Muhammet Fatih, Akif Durdu, and Kadir Sabanci. "Human action recognition with a bag of visual words using different machine learning methods and hyperparameter optimization." Neural Computing and Applications (2019): 1-13.

[2] Al-Akam, Rawya, and Dietrich Paulus. "Local Feature Extraction from RGB and Depth Videos for Human

Action Recognition." International Journal of Machine Learning and Computing 8, no. 3 (2018).

[3] Al-Akam, Rawya, and Dietrich Paulus. "RGBD Human Action Recognition using Multi-Features Combination and K-Nearest Neighbors Classification." International Journal of Advanced Computer Science and Applications (IJACSA) 8, no. 10 (2017): 383-389.

[4] Matsufuji, Akihiro, Wei-Fen Hsieh, Hao-Ming Hung, Eri Shimokawara, Toru Yamaguchi, and Lieu-Hen Chen. "A Method of Action Recognition in Ego-Centric Videos by Using Object-Hand Relations." In 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI), pp. 54-59. IEEE, 2018.

[5] Nazir, Saima, Muhammad Haroon Yousaf, Jean-Christophe Nebel, and Sergio A. Velastin. "Dynamic Spatio-Temporal Bag of Expressions (D-STBoE) model for human action recognition." Sensors 19, no. 12 (2019): 2790.

[6] Mishra, Om, Rajiv Kapoor, and M. M. Tripathi. "Human Action Recognition Using Modified Bag of Visual Word based on Spectral Perception." (2019).

[7] Najar, Fatma, Sami Bourouis, Nizar Bouguila, and Safya Belghith. "Unsupervised learning of finite full covariance multivariate generalized Gaussian mixture models for human activity recognition." Multimedia Tools and Applications (2019): 1-23.

[8] Aly, Saleh. "An effective human action recognition system based on Zernike moment features." In 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), pp. 52-57. IEEE, 2019.

[9] Naveed, Humza, Gulraiz Khan, Asad Ullah Khan, Aiman Siddiqi, and Muhammad Usman Ghani Khan. "Human activity recognition using mixture of heterogeneous features and sequential minimal optimization." International Journal of Machine Learning and Cybernetics 10, no. 9 (2019): 2329-2340.

[10] Cortés, Xavier, Donatello Conte, and Hubert Cardot. "A new bag of visual words encoding method for human action recognition." In 2018 24th International Conference on Pattern Recognition (ICPR), pp. 2480-2485. IEEE, 2018.

[11] Avola, Danilo, Marco Bernardi, and Gian Luca Foresti. "Fusing depth and colour information for human action recognition." Multimedia Tools and Applications 78, no. 5 (2019): 5919-5939.

[12] Avola, Danilo, Marco Bernardi, and Gian Luca Foresti. "Fusing depth and colour information for human action recognition." Multimedia Tools and Applications 78, no. 5 (2019): 5919-5939.

[13] Sun, Guangmin, Chenyang Wang, Beichuan Ma, and Xiaomeng Wang. "An improved SIFT algorithm for infringement retrieval." Multimedia Tools and Applications 77, no. 12 (2018): 14745-14765.

[14] Liu, Yijian, King-Chi Fung, Wenqian Ding, Hongfei Guo, Ting Qu, and Cong Xiao. "Novel Smart Waste Sorting System based on Image Processing Algorithms: SURF-BoW and Multi-class SVM." Computer and Information Science 11, no. 3 (2018): 35-49.

[15] Plötz, Thomas, and Yu Guan. "Deep learning for human activity recognition in mobile computing." Computer 51, no. 5 (2018): 50-59.

[16] Nour el houda Slimani, Khadidja, Yannick Benezeth, and Feriel Souami. "Human interaction recognition based on the co-occurence of visual words." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 455-460. 2014.

[17] Liu, Hong, Hao Tang, Wei Xiao, ZiYi Guo, Lu Tian, and Yuan Gao. "Sequential Bag-of-Words model for human action classification." CAAI Transactions on Intelligence Technology 1, no. 2 (2016): 125-136.

[18] Jiang, Xinghao, Tanfeng Sun, Bing Feng, and Chengming Jiang. "A space-time surf descriptor and its application to action recognition with video words." In 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), vol. 3, pp. 1911-1915. IEEE, 2011.

[19] Xu, Ke, Xinghao Jiang, and Tanfeng Sun. "Human activity recognition based on pose points selection." In 2015 IEEE International Conference on Image Processing (ICIP), pp. 2930-2834. IEEE, 2015.

[20] Megrhi, Sameh, Marwa Jmal, Wided Souidene, and Azeddine Beghdadi. "Spatio-temporal action localization and detection for human action recognition in big dataset." Journal of visual communication and image representation 41 (2016): 375-390.

[21] Sabri, Aznul Qalid Md, Jacques Boonaert, Stéphane Lecoeuche, and El Mustapha Mouaddib. "Multi Spatio-Temporal Co-occurrence Measures for Human Action Classification." In MVA, pp. 319-322. 2013.

[22] Megrhi, Sameh, Azeddine Beghdadi, and Wided Souideǹ e. "Trajectory feature fusion for human action recognition." In 2014 5th European Workshop on Visual Information Processing (EUVIP), pp. 1-6. IEEE, 2014.

[23] Zhang, Jia-Tao, Ah-Chung Tsoi, and Sio-Long Lo. "Scale invariant feature transform flow trajectory approach with applications to human action recognition." In 2014 International Joint Conference on Neural Networks (IJCNN), pp. 1197-1204. IEEE, 2014.

[24] Harjanto, Fredro, Zhiyong Wang, Shiyang Lu, Ah Chung Tsoi, and David Dagan Feng. "Investigating the impact of frame rate towards robust human action recognition." Signal Processing 124 (2016): 220-232.

[25] Malpani, Sourabh, C. S. Asha, and A. V. Narasimhadhan. "Thermal vision human classification and localization using bag of visual word." In 2016

IEEE Region 10 Conference (TENCON), pp. 3135-3139. IEEE, 2016.

[26] Sabri, Aznul Qalid Md, Jacques Boonaert, Erma Rahayu Mohd Faizal Abdullah, and Ali Mohammed Mansoor. "Spatio-Temporal Co-Occurrence Characterizations For Human Action Classification." arXiv preprint arXiv:1610.05174 (2016).

[27] Chattopadhyay, Chiranjoy, and Sukhendu Das. "Supervised framework for automatic recognition and retrieval of interaction: a framework for classification and retrieving videos with similar human interactions." IET Computer Vision 10, no. 3 (2016): 220-227.

[28] Avgerinakis, Konstantinos, Alexia Briassouli, and Ioannis Kompatsiaris. "Activities of daily living recognition using Optical trajectories from motion boundaries." Journal of Ambient Intelligence and smart environments 7, no. 6 (2015): 817-834.

[29] Ahad, Md Atiqur Rahman. "Action Representation Approaches." In Computer Vision and Action Recognition, pp. 39-76. Atlantis Press, 2011.

[30] Baccouche, Moez, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. "Sequential deep learning for human action recognition." In International workshop on human behavior understanding, pp. 29-39. Springer, Berlin, Heidelberg, 2011.

[31] M. S. RYOO AND J. K. AGGARWAL, UT-interaction dataset, (2010).

[32] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." In European conference on computer vision, pp. 404-417. Springer, Berlin, Heidelberg, 2006.