Human Action Recognition in Video Sequence using Logistic Regression by Features Fusion Approach based on CNN Features

Tariq Ahmad[1], Imran Khan[3]
School of Information and Communication Engineering
Guilin University of Electronic Technology
Guilin, China

Asif Rahim[4]
School of Cyberspace Security
Guilin University of Electronic Technology
Guilin, China

Jinsong Wu[2]*
School of Artificial Intelligence
Guilin University of Electronic Technology
Guilin, China

Amjad Khan[5]
Hamdard Institute of Engineering & Technology
Hamdard University, Islamabad Campus
Islamabad, Pakistan

***Abstract*—Human Action recognition (HAR) gains too much attention due to its wide range of real world applications, such as video surveillance, robotics and computer vision. In video surveillance systems security cameras are placed to monitor activities and motion, generate alerts in undesirable situations. Due to such importance of video surveillance in daily life, HAR becomes the primary and key factor of video surveillance systems. Many researchers worked on human action recognition but HAR still a challenging problem, due to large variation among human to human and human actions in daily life, which make human recognition very challenging and makes surveillance system difficult to outperform. In this article a novel method is proposed by features fusion of pre-trained convolution neural network (CNN) features. Initially pre-trained CNN VGG 19 weights are exploited to extract fully connected 7th layer (FC7) of the selected dataset, subsequently pre-trained fully connected 8th layer features (FC8) extracted by employing pre-trained weights of the same neural network. However the resultant feature fused vector further optimized by employing two statistical features selection techniques, chi-square test and mutual information to select best features among them to reduced redundancy and increase performance accuracy of human action, a threshold value used for selecting best features. Furthermore the best features are fused, then grid search with 10 fold cross validation is applied for tuning hyper parameter to select best k fold and the resulting best parameter are feed to Logistic regression (LR) classifier for recognition. The proposed technique used You Tube 11 action dataset and achieved 98.49% accuracy. Lastly the proposed method compares with the existing state of the art methods which show dominance performance.**

Keywords*—Human action recognition; logistic regression; deep learning; convolution neural network; features fusion

I. Introduction

Human action recognition (HAR) is very popular among researchers, computer vision community, data engineers and data scientist due to its wide range of industrial and real life applications. One of the main inspirations which invite scholars to work in human action recognition is the wide domain of its applications in computer vision, robotics, human computer interaction [1], and video surveillance [2], in the former HAR technique faced challenges due to similarity of visual contents whereas the later one faced challenges due to large variation among humans in real life. In video surveillance systems security cameras installed to monitor activities of human and generate alerts in undesirable situation, store videos and transmitting videos but due to huge variation among intrapersonal and personal activities of humans in real life make video surveillance systems difficult to outperform because human action recognition is the basic feature which directly impacts the performance of video surveillance system.

Human action is the motion of body portions by interacting with components in environment. Human action may be simple like movement of arm or leg e.g. walking activity of human includes arm and leg movements, and may be too complex like movement of entire body e.g. jumping of volleyball player, which includes movements of entire body of the player. HAR methods are used in wireless sensor networks [3], wearable sensor [4] and video HAR [5] but HAR is more popular in video based systems because video based HAR techniques are the basic building block of video surveillance system and video surveillance systems are extensively used in real life.

Human action recognition in videos sequence is the process of allotting labels to each category of videos to train the system, and the system enable to recognize various actions done by human in unseen videos, however in the context of videos an action is embodied using sequences of frames from which humans can easily understand by examining contents of numerous frames in sequence [6]. Human action recognition in videos is still challenging due to many factors such as class variation, angle variation [7] and environment. Researchers used many techniques for image classification and action recognition such as hand-crafted methods [8], and Histogram of Oriented Gradient (HOG), but such hand-crafted techniques have some limitations such as complex computations and lengthy videos which run continuously, however hand-crafted

*Corresponding Author.

techniques outperform in some domain of action recognition such as simple videos streaming.

In recent decade researcher also used deep learning networks methods for many applications of action recognition and image classification, over millions of multi-class images classified through CNN based approaches, which proved the accuracy is improved in classification problem [9], [10]. CNN based approaches shown significant improvement in many areas which give them too much consideration over hand-crafted features techniques. Many researchers presented their work for action recognition by using pre-trained weights of deep learning neural networks for features extraction, instead of training new deep network from scratch. For instant, building a new model from scratch need huge amount of data which is computationally expensive, relatively, on small dataset the deep learning network from scratch will be not outperform due to small amount of data.

In this paper, we used six frames per second of each video clip instead of thirty frames per second; we bounce five frames every second of every video clip, which reduce redundancy and computation complexity. In the proposed method weights of pre-trained CNN [9], used for feature extraction and then feed these deep features to logistic regression for action recognition of sequence frames of selected video dataset. The pre-trained CNN model was selected on the basis of its prior performance over classification problem and due to few limitations of hand-crafted features based methods we selected deep neural network based approach for feature extraction in the proposed research method. A detail overview presented in subsequent sections.

II. Related Work

Over the years researchers presented their work for action recognition, based on hand-crafted and deep learning networks. Both techniques discuss in Section A and B respectively.

A. Hand-crafted based Features Extraction Techniques

Hand-crafted based approaches extract hand-crafted features from simple video clip for non-realistic actions, where a performer completes an action in a scene with simple context and situation; hand–crafted techniques extract low level feature map of human action in video sequence and feed these features to classifier such as ensemble, naive Bayes and (SVM) support vector machine for action recognition. In [11], action sketches were investigated by analysis of geometric characteristic such as space, time and volume (STV). In [12,] the author presented human action as three dimensions prepared from silhouettes in space, time and volume, moreover Poisson’s equation used by them to examine two dimensions shape of actions and exposed space time features (STF) for non-realistic videos sequence, however two dimensions shape of actions for two different actions sometime caused the same shape and making the action recognition difficult. In [13], the author used realistic video dataset and extracted motion and static features; in addition they removed the noisy features by applying motion statistics and obtain stable features. However hand-crafted features techniques have certain drawbacks. For instant, STV based approach are not effective for recognition of numerous person activities in a scene. STF based approaches are not appropriate for complex dataset however its shows significant result on simple dataset. These drawbacks can cause trouble for lengthy videos and real time applications with nonstop video streaming such as video surveillance systems.

B. Deep Neural Networks Techniques

In recent years several deep learning networks for action recognition, image classification, bioinformatics and person re-identification were presented and show significant accuracy in the respected fields. For instant, a straight forward execution were developed for human activity recognition [14], moreover they used 3D CNN filters in implementation and applied on videos frames in time domain to capture spatial and temporal information. They also claimed that their proposed technique collects optical and motion features, since video frames were linked to fully connected layer at the end of deep network.

A multi-resolution convolution neural network was presented in [15], to collect local spatial and temporal features, they used time axis for connectivity of features. They tested the experiment on YouTube one million videos dataset for human action recognition and acquired 63.9% recognition, they claimed that the proposed work reduce time complexity in training the system, but their recognition is still low for other large action recognition dataset such as UCF101[16] recognition was 63.3%.

Two stream convolution neural networks was presented in [17], to captured spatial and temporal features in video frames they used first stream, moreover the second stream captured optical flow of frames in dense. Asymmetric unidirectional 3D CNN was presented [18], for recognition they applied micro nets to increase feature learning skill of their proposed deep learning network and achieved good recognition rate. Deep learning based techniques have the capability to correctly detect unseen patterns in visual data because of its vast quantity of data for training and huge computational power for its processing.

III. Callanges and Contributions

In the recent decade significant contribution was made in human action recognition (HAR). Many approaches were applied in HAR aimed to acquire high recognition of human actions in the domain of HAR, but mostly hand-crafted based and deep learning based methods were presented in literature by researchers. In hand-crafted based approaches such as STV, action sketches were examined by geometric characteristics in time, space and volume but STV based approach was only effective for single actor action where an actor perform some action in the scene, STV got optimal results, conversely STV based approach was behind to solve the challenge of HAR where multiple actors perform actions in the scene. Some hand-crafted based approaches such as space, time features STF used human silhouettes to examine two dimensions shape of actions and expose space time features but STF was slow and consume huge amount of space.

Which deviate time and space tradeoff, in STF the two dimensions shape of action for two different actions sometime produced the same human action which caused difficulty in recognition and the final accuracy of the system affected. Beside hand-crafted approaches deep neural networks made significant contribution towards solutions to the challenges faced by HAR, deep neural networks show significant results in many domains, In addition deep neural networks based approaches relies on very deep features and many real life applications which used HAR are also rely on deep features. However, an end to end HAR system depends on very deep features for outstanding results therefore in the proposed research deep neural networks based features extraction techniques is used. In this paper the main contributions towards the solutions to HAR in video sequence are:

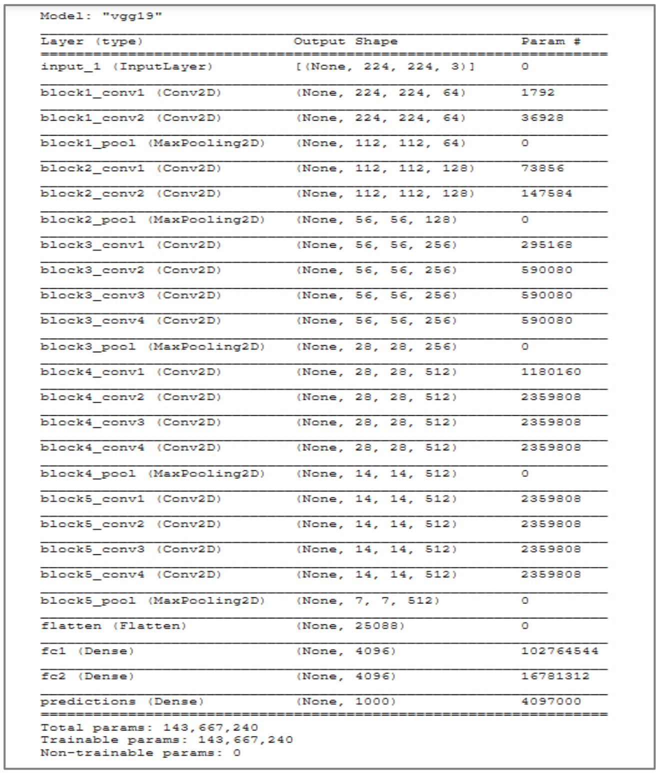
- Utilizing VGG19 model to compute deep features and acquire two features vectors of our selected video dataset.
- Integrated the deep features of CNN and computing best features by applying Chi-2 and mutual information to reduce redundancy and increase performance.
- Integrated Chi-2 best features and mutual information features vector and apply grid search with 10 k-fold cross validation to tune hyper parameter and select best k-fold parameters.
- Finally, the resulting k-fold is feed to LR for final recognition to solve recognition problem.

IV. Proposed Research Methodology

In this section the proposed research methodology is discusses in details. An activity A_c in frames of video sequence V_d using CNN for features extraction and logistic regression (LR) for F_R sequence of frames to recognize $A_{c.}$ Firstly we use pre-trained weights of CNN for feature extraction of frames F_R in video V_d with bounce of B_f such that bouncing of frames not affect the activity A_c. finally the features fed to logistic regression for activity A_c recognition.

A. Preprocessing of Input Frames of Video Sequence

Video is the collection of frames generally video is running at thirty frames per second but we take into consideration only six frames per second and bounce five frames at unit time which reduced redundancy and computational complexity, however the selected sequence of frames doesn't affect action in video and from the evaluation of experiment it achieved significant result. In the preprocessing phase we resize frames to 224x224 RGB of all categories which is the desire input shape of VGG 19 [9] for feature extraction. In the context of videos, frames are features of videos so every frame is zero centered to reduced computation and preprocessed all frames to subtract mean RGB pixel intensity from pre-trained weights of VGG 19 during feature extraction phase. VGG19 model trained on 1.3 million images and 143 million trainable parameters which allow VGG19 model to transfer the learned pattern from pre-trained weights to our selected dataset in feature extraction. The architecture of VGG19 model is given in Fig. 1.



```
Model: "vgg19"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 224, 224, 3)]     0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 224, 224, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 224, 224, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 112, 112, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 112, 112, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 112, 112, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 56, 56, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 56, 56, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 56, 56, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 56, 56, 256)       590080
_________________________________________________________________
block3_conv4 (Conv2D)        (None, 56, 56, 256)       590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 28, 28, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 28, 28, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 28, 28, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 28, 28, 512)       2359808
_________________________________________________________________
block4_conv4 (Conv2D)        (None, 28, 28, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 14, 14, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_conv4 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 7, 7, 512)         0
_________________________________________________________________
flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
fc1 (Dense)                  (None, 4096)              102764544
_________________________________________________________________
fc2 (Dense)                  (None, 4096)              16781312
_________________________________________________________________
predictions (Dense)          (None, 1000)              4097000
=================================================================
Total params: 143,667,240
Trainable params: 143,667,240
Non-trainable params: 0
```

Fig. 1. VGG19 Model Architecture.

VGG19 model used fix input shape of RGB images in the training phase, they used 1.3 million images for training the model, 50 thousand for validation and 100 K images for evaluation the experiment, stack of convolution layers conv1 and conv2 were employed to input RGB image with 3x3 filter size to extract low level features of images, passing the input RGB images from conv1 and conv2 the image RGB channel is converted to 64, in the first block of convolution layers, stride of 1 pixel used for sliding the filter map and max pool1 layer to reduce spatial size of conv1 and conv2 features. In the second block of convolution layers 3x3 conv1 and conv2 applied followed by max pool layer.

They used stride of 1 pixel for every convolution and 2x2 strides for max pooling layers; however activation function of ReLU [19], equipped in all hidden layers for rectification and introduced non-linearity form which the model learned complex useful features between inputs and response variables. Beside stack of convolution layers one flatten layer of 250,88 dimension applied and then followed by three fully connected layers FC1, FC2 each have 4096 channels depth and the last FC layer of 1000 channels, finally a soft-max function used for prediction of classes. The model trained on 143 million parameters, initial learning rate of 10^{-2}, momentum 0.9, number of iteration 370 K and mini batch-size of 256 used respectively. The proposed research methodology is given in Fig. 2. Where each step is discuss in subsequent sections.

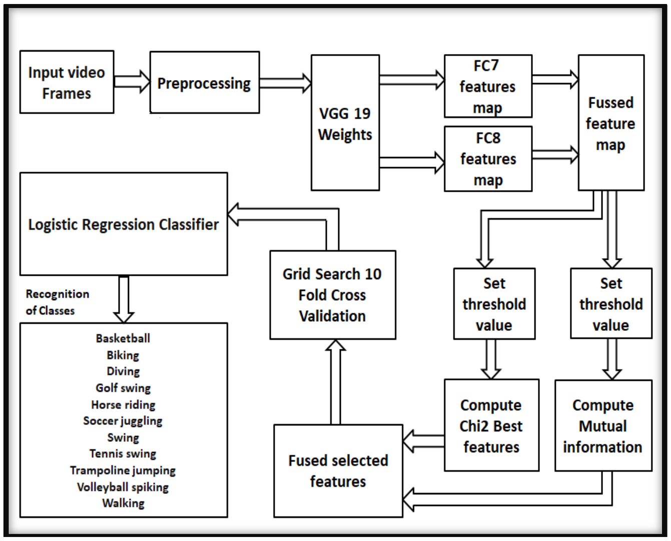


Fig. 2. Basic Building Block of Proposed Framework.

B. Features Extraction and Features Fusion

Video clip is the collection of frames where its running 30 frames at unit time by default but we used six frames per second and bounce five frames in each second because frames represents the story of the running video and at rate of 30 frames per unit time cause a very small movement in the story and cause redundancy, from which the system computational complexity increased, in Fig. 3. Scenario of a sample video clip is presented whereas frames are extracted in one second and frame to frame change in unit time occurs during frame extraction. We passed frames of our selected dataset through VGG19 weights and extracted features from fully connected 7th and 8th layer respectively. If we represent the extracted features in vector representation of *N* data samples and *d* dimensions then the extracted features vectors can be denoted (N, d_1) and (N, d_2) where d_1 and d_2 represent the features dimensions of FC7 and FC8, respectively. In addition, the extracted features vectors can be express in equation such that

$$V^{(7)} = (N, d_1) \tag{1}$$

$$V^{(8)} = (N, d_2) \tag{2}$$

Where $V^{(7)}$ and $V^{(8)}$ represents extracted feature vectors of 7th and 8th layers of VGG19, respectively.

The given Fig. 4 shows the visualization effect of filters after applying VGG 19 activation function on the sample frame of selected dataset during features extraction. furthermore, in the proposed method, we used features fusion technique by applying vector addition to (1) and (2) and acquire fused vector V_f, however (1) and (2) used the same sample of data then in vector addition the size of N taken common.

Equation (1) and (2) by transformation of vector addition.

$$V_f = V^{(7)} + V^{(8)} \tag{3}$$

$$V_f = (N, d_1) + (N, d_2) \tag{4}$$

$$V_f = (N+N, d_1+d_2) \tag{5}$$

$$V_f = (N_i, d_n) \tag{6}$$

Where $N_i = N$ such that $N_i \neq N+N$ and $d_n = d_1+d_2$

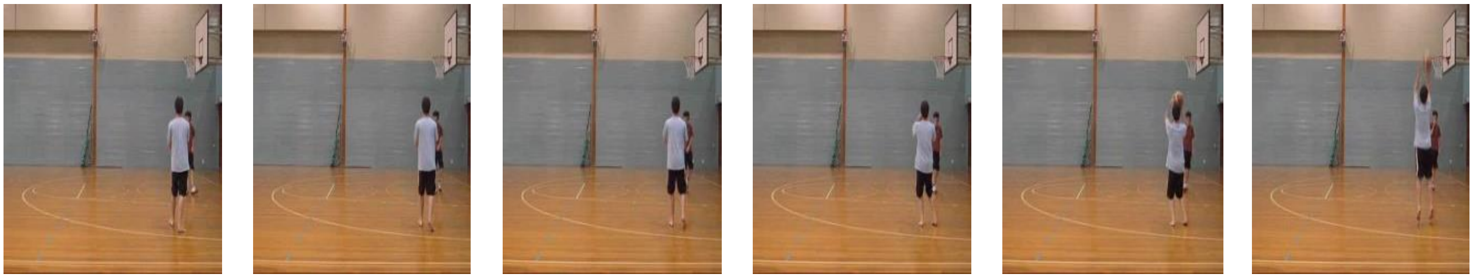


Fig. 3. Frame to Frame Representation and Change in Frames Occurs in One Second.

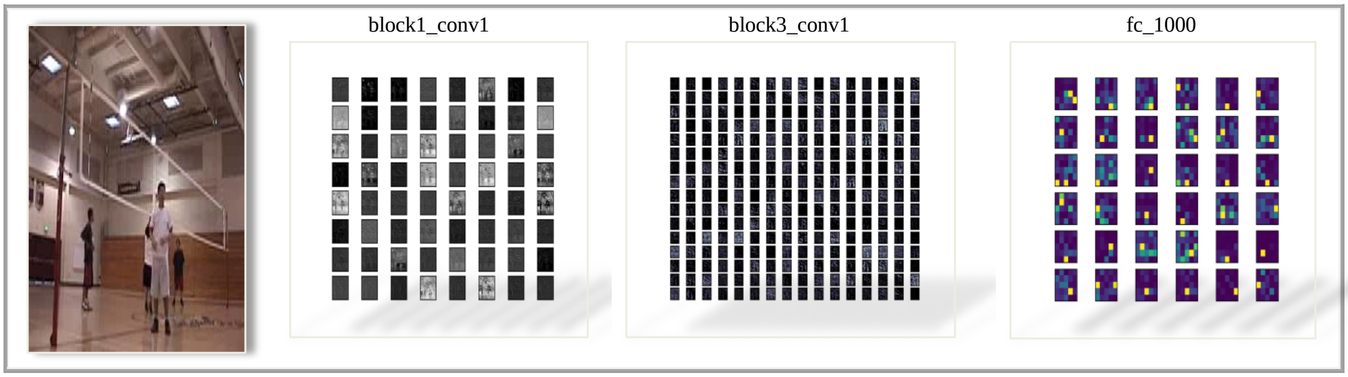


Fig. 4. Sample Frame of You Tube Dataset and Visualization of Filters after Applying Activation Function.

C. Features Selection

Feature selection is very useful when building a machine learning model because not all features in dataset are useful and adding all features to model may reduce the accuracy of model and generalization capacity. Furthermore, model complexity also increases with increasing the number of useless features. We have fused features vector V_f which contains redundant features, to decrease redundancy from features and select best features we apply two statistical feature selection techniques, chi-square test and mutual information on V_f. chi-square test compute between features and target variable and select best features based on chi-square scores, for instant if some features get low chi-square score then remove those features we applied a threshold value in our proposed method. The mathematical representation of chi-square:

$$X_c^2 = \frac{\sum (\mathrm{O}-\mathrm{E})^2}{\mathrm{E}} \qquad (7)$$

Where

C = degree of freedom

O = observed values

E = expected values

We have fused features vector V_f and a matrix L_n where L_n represents class labels of training samples, however in chi-square features selection technique L_n consider as target variables of n dimension then by putting V_f and L_n in (7) we can get.

$$X_c^2 = \frac{\sum (\mathrm{Vf}-\mathrm{L}_n)^2}{\mathrm{L}_n} \qquad (8)$$

Here X_c^2 contains the score of each feature acquire from chi-square, moreover X_c^2 scores are applied to transformed fused features vector under certain threshold such that

$$\mathrm{Ch_b}=\mathrm{X_c}^2\mathrm{T}\rightarrow\mathrm{V_f} \qquad (9)$$

Where ch_b represents best chi-square features, → denote transformation function and T is threshold value respectively. Beside chi-square feature selection, mutual information feature selection technique is also computed in the proposed research, mutual information between two variables is the measurement that how much information obtains one variable through the other variable. Mathematically formulation of mutual information is;

$$\mathrm{MI}\ (\mathrm{A};\mathrm{B}) \vartriangle \ \mathrm{D}\ (\mathrm{P_{AB}} \Vert \mathrm{P_A P_B}) \qquad (10)$$

Where A and B are independent variables, P_{AB} is joint probability density function of A and B, where P_A and P_B are marginal density function of variables A and B respectively. Consequently, (10) applied on V_f and L_n.

$$\mathrm{MI}\ (\mathrm{V_f}\,;\mathrm{L_n}) \vartriangle \ \mathrm{D}\ (\mathrm{P}_{\mathrm{Vf}} \Vert \mathrm{P}_{\mathrm{Vf}}\, \mathrm{P_{Ln}}) \qquad (11)$$

$$\mathrm{MI_h} = \mathrm{MI}\ (\mathrm{V_f}\,;\mathrm{L_n}) \vartriangle \ \mathrm{D}\ (\mathrm{P_{AB}} \Vert \mathrm{P_A P_B})\ \mathrm{T} \rightarrow \mathrm{V_f} \qquad (12)$$

Where MI_h contains only those features which have high mutual information between V_f and L_n under certain threshold, → used for transformation function and T denote threshold respectively. Besides Ch_b and MI_h, whereas the former holds best score chi-square features and the later one contains high mutual information features, we fused both the vectors through vector addition.

$$\mathrm{V_S} = \mathrm{Ch_b} + \mathrm{MI_h} \qquad (13)$$

Where V_S denote selected features.

D. Logistic Regression

Logistic Regression (LR) widely used in many applications of data mining and machine learning techniques for data classification. LR delivers likelihoods and cover multi-class classification problem [20]. LR used the same principle of linear regression, furthermore LR techniques were applied through truncated newton to solve large optimization problem [21]. However LR applied in many imbalance and multi-class data to solve the classification problem. We can express LR mathematically as.

$$P = \frac{e^{a+bx}}{1+e^{-(b_0+b_1x)}} \qquad (14)$$

Where we have fused selected vector V_S by feeding this vector to (14) and solve the classification problem for action recognition. In addition, prior to feeding V_S to (14) we choose a set of optimal hyper-parameter by grid search to get best fit of proposed LR model, and further we applied 10 k-fold cross validation to evaluate the proposed LR model.

For instant, we have fused selected features V_S which contains multi-class features and imbalance data samples because V_S fused features of videos frame and in context of videos, not all videos are same in size, some may be lengthy, short and medium in size. To avoid the imbalance data problem and balance all categories we gave equal class weights to every categories of selected dataset, because imbalance data directly impacts on average accuracy of machine learning model. Fig. 5 shows the imbalance frames for our selected dataset.

V. Experimental Evaluation

In this section we will discuss the dataset and results of the proposed research methodology which based on pre-trained CNN features and select best features by using Chi-2 and mutual information. The experiment is evaluated on publically available benchmark You Tube 11 action dataset; first we tested our proposed method on three classifiers Logistic Regression (LR), Naïve Bayes (NB) and Random Forest (RF) and then chose best classifier among them based on performance result. We chose LR because LR achieved significant results. Table I show the comparison results of LR, NB and RF. Next, the proposed method using LR is compared with some of the existing state-of- the art techniques of HAR. Initially the dataset divided into 80% for training and 20% for testing, according to machine learning standard protocol for data splitting. In the proposed research method we used deep learning framework, tensorflow for deep feature extraction, for features fusion and implementation of LR model we used python Sklearn library. Overall experiment tested on NVidia GTX 1080ti GPU and 16 GB of RAM used respectively.

A. You Tube Action Dataset

We tested and evaluate our proposed method using LR on you tube action dataset which is publically available, it contains 11 action classes: basketball, biking, diving, golf, horse riding, soccer, swing, tennis, trampoline jumping, volleyball, and walking. The dataset contains 11 classes and each class provided a subset of 25 groups further, whereas every group of videos contains more than 4 videos clips, however the videos in same group share some similarity such as identical actors, background of video clips matched to other videos of the same group and equal viewpoint [27]. The given Fig. 6 shows some frames sample which represent different categories of You Tube 11 action dataset.

The dataset is very challenging due to pose and object appearance, cluttered background, large variation in camera motion, viewpoint, illumination condition and object scale, some videos of the same class captured when an actor done some actions in indoor background where others videos of the same class captured while an action done by actor some in outdoor background.

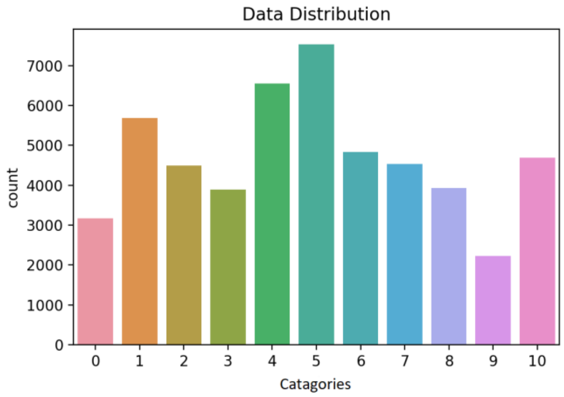
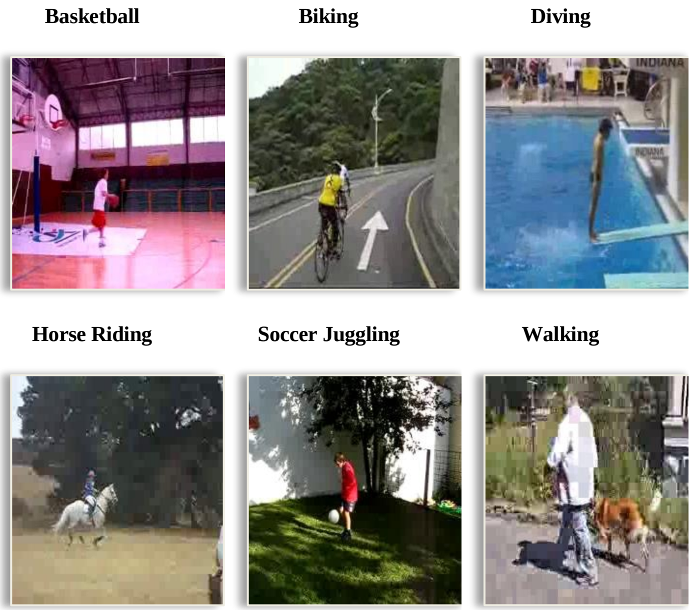


Fig. 5. Imbalance Frames of Selected Video Dataset.


Basketball

Biking

Diving

Horse Riding

Soccer Juggling

Walking

Fig. 6. Different Samples Frames of You Tube Dataset.

B. *Selection of Classifier*

Selection of classifier to get optimal results to solve recognition problem faced by HAR, we tested our proposed method of features extraction and selection on three different classifier LR, NB and RF using you tube 11 dataset, whereas the feature are extracted through VGG19 model and select best features through Chi-2 and mutual information, for all the three classifier same methodology used for features extraction and features selection, the aim of testing the proposed method on three classifiers is to select best classifier among them and to check the consistency and validity of our results. The given Table I reported the recognition accuracy, F1 score, recall and precision of all the three classifiers. We choose LR to solve recognition problem faced by HAR because LR achieved 98.55% F1 score, 98.57% recall and 98.53% precision where NB achieved 78.41% F1 score, 80.71% recall and 77.97% precision and FR achieved 93.22% F1 score, 92.55% recall and 94.06% precision by using the selected dataset respectively. The given Fig. 7 shows comparison results of LR, NB and RF on test samples of YouTube dataset by using the proposed method of feature extraction and selection. LR classifier got optimal results over Naïve Bayes and Random forest; the former achieved 79.50% accuracy whereas the later one achieved 93.23% and selected LR achieved 98.49% recognition accuracy on test data. The selection of LR in the proposed research is not only based on accuracy, but we used total four metrics for the selection criteria of classifier, further, from evaluation results of LR, NB and RF which exploit our proposed method of features extraction and selection shows the validity and consistency in results.

TABLE I. NB, RF AND LR RESULTS COMPARISON

Method	Accuracy	F1 score	Recall	Precision
Navie Bayes	79.50%	78.41%	80.71%	77.97%
Random forest	93.23%	93.22%	92.55%	94.06%
Logistic Regression	98.49%	98.55%	98.57%	98.53%

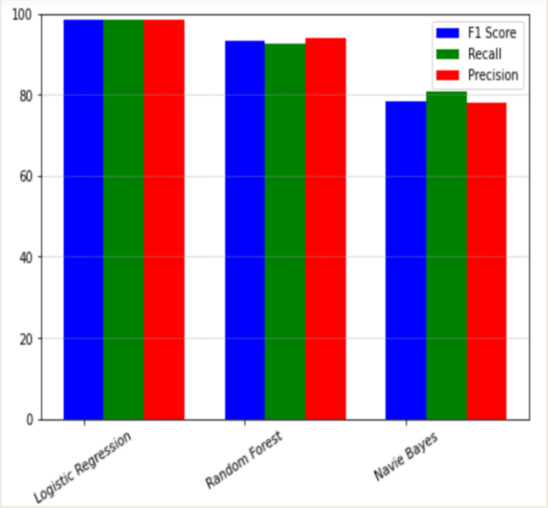


Fig. 7. Comparison of LR, NB and RF Tested on You Tube Action Dataset.

C. Proposed LR Comparison with Existing Techniques

The proposed method using LR achieved 98.49% average accuracy on test samples of You Tube action dataset given in Table II, dominating the Wang, Wu, Yang, Xu, Peng techniques having 84.1%, 87.0%, 88.0%, 89.3%, 93.8% accuracy, respectively. The confusion matrix of You Tube action dataset evaluated on test samples is given in Fig. 8. The proposed LR method achieved more than 98% accuracy for seven classes among eleven. The class "walking" reported 96.7% accuracy because some other classes interfere and reported false prediction of 3.3%, similarly class "biking", "swing" and "trampoline jumping" accuracy are 97.7%, 97.7% and 97.8% reported respectively because other classes intervene due to same view point, background etc. and affect average recognition accuracy. Fig. 9 shows class wise accuracy of You Tube action recognition dataset which are evaluated on test data. Our proposed method using LR achieved significant results and by comparison with existing state of the art techniques we conclude that our method is best fit for solving the recognition problem of HAR.

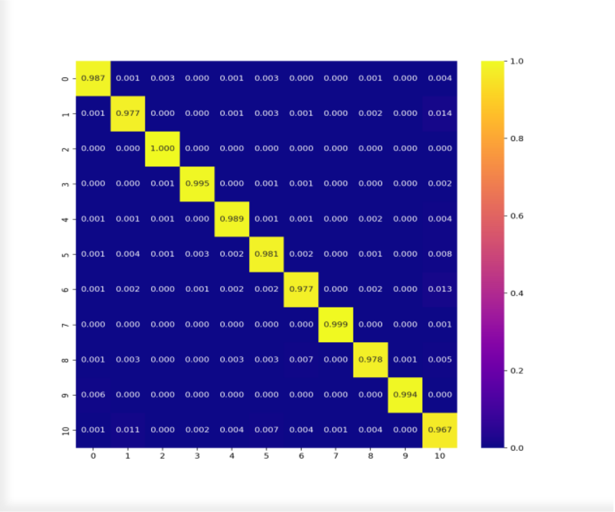


Fig. 8. Confusion Matrix of You Tube Dataset using LR.

TABLE II. COMPARISON OF AVERAGE ACCURACY OF PROPOSED METHOD FOR ACTION RECOGNITION WITH STATE-OF THE-ART TECHNIQUES

Method	**You Tube Dataset**
Wang[22]	84.1%
Wu[23]	87.0%
Yang[24]	88.0%
Xu[25]	89.3%
Peng[26]	93.8%
Proposed	98.49%

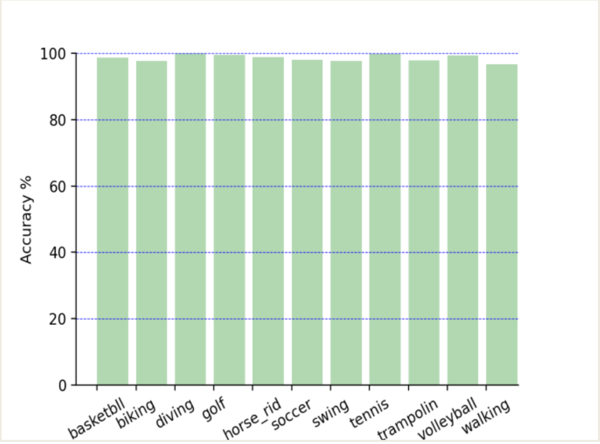


Fig. 9. Class wise Accuracy of Test Samples of You Tube Dataset.

VI. CONCLUSION AND FUTURE WORK

Human action recognition (HAR) under many viewpoints is a major challenge to correctly recognize the activity of human. In this work we proposed a new approach for HAR. First we extracted deep features from fully connected 7th (FC7) and 8th (FC8) layers of pre-trained model namely VGG19, and next, integrate the two features vector to select best features among them by applying Chi-2 and mutual information, later Logistic Regression used for classification by feeding the best features acquired from Chi-2 and mutual information. The aim of selection of best features is to improve accuracy and reduce redundancy from features, however feeding noisy features to the system must be consume more time and make the system computation expensive. The experiments are conducted on You tube 11 action dataset and the proposed method outperform, from the experiment we concluded that features extraction from pre-trained model perform better for improvement of recognition. Our method achieved 98.49% average accuracy on you tube 11 dataset and by comparison with existing state-of-the-art-techniques our method dominating in performance. In future, we are planning to use some advance dataset UCF 50 and UCF101 which contains 50 and 100 categories of human actions respectively. Furthermore we are planning to use gaited recurrent unit (GRU) to solve the recognition problem of HAR.

ACKNOWLEDGMENT

This work is supported in part by the School of Information and Communication Engineering, Guilin University of Electronic Technology, the School of Artificial Intelligence, Guilin University of Electronic Technology, the National Natural Science Foundation of China (No's, 6217011456, 6216010122).

REFERENCES

[1] S. A. Aly, T. A. Alghamdi, M. Salim, and A. A. Gutub, ''Data dissemination and collection algorithms for collaborative sensor devices using dynamic cluster heads,'' *Trends Appl. Sci. Res.*, vol. 8, no. 2, pp. 55–72, 2013.

[2] A. Nanda, P. K. Sa, S. K. Choudhury, S. Bakshi, and B. Majhi, ''A neuro- morphic person re-identification framework for video surveillance,'' *IEEE Access*, vol. 5, pp. 6471–6482, 2017.

[3] R. Zhao, W. Xu, H. Su, Q. Ji, "Bayesian Hierarchical Dynamic Model for Human Action Recognition," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7733-7742.

[4] S. Rahimi, A. Aghagolzadeh, M. Ezoji , "Human action recognition based on the Grassmann multi-graph embedding," Signal, Image and Video Processing volume 13, pages 271–279, 2019.

[5] A-A. Liu, Y-T. Su, W-Z. Nie, M. Kankanhalli, "Hierarchical Clustering Multi-task Learning for Joint Human Action Grouping and Recognition," IEEE Transaction on pattern anaylisis and machine intelligence, DOI:10.1109/TPAMI.2016.2537337, February 2016.

[6] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, N. Zheng, "View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, PP. 99, DOI:10.1109/TPAMI.2019.2896631.

[7] Z. Tu, W. Xie, Qi. Qin, R. Poppe, R. C. Veltkamp, B. Li, J. Yuan, "Multi-stream CNN: Learning representations based on human-related regions for action recognition" *Pattern Recognition* (IF7.74), Pub Date : 2018-07-01, *DOI: 10.1016/j.patcog.2018.01.020*.

[8] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, ''Action detection in complex scenes with spatial and temporal ambiguities,'' in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 128–135.

[9] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition". ICLR 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classifica - tion with deep convolutional neural networks,'' in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[11] A. Yilmaz and M. Shah, ''Actions sketch: A novel action representation,'' in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2005, pp. 984–989.

[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, ''Actions as space-time shapes,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. no. 12, pp. 2247–2253, Dec. 2007. 29.

[13] J. Liu, J. Luo, and M. Shah, ''Recognizing realistic actions from videos 'in the wild,''' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1996–2003.

[14] S. Ji, W. Xu, M. Yang, and K. Yu, ''3D convolutional neural networks for human action recognition,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, ''Large-scale video classification with convolutional neural networks,''in*Proc.IEEEConf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[16] K. Soomro, A. R. Zamir, and M. Shah. (2012). ''UCF101: A dataset of 101 human actions classes from videos in the wild.'' arXiv preprint arXiv:1212.0402, 2012.

[17] K. Simonyan and A. Zisserman, ''Two-stream convolutional networks for action recognition in videos,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[18] H. Yang, C.Yuan, B .Li, Y .Du, J. W .Xing, Hu et al, (2019) "Asymmetric 3d convolutional neural networks foraction recognition". Pattern Recogn 85:1–12.

[19] A. Fred M.Agarap "Deep Learningusing RectifiedL inear Units (ReLU," [online]. Available: https://arxiv.org/pdf/1803.08375.pdf.

[20] T. Hastie, R. Tibshirani, and J. Friedman, (2009), "The Elements of Statistical Learning", 2nd ed., Springer Verlag.

[21] P, Komarek, and Moore, A. (2005b), "Making logistic regression a core data mining tool with TR-IRLS", *Proceedings of the Fifth IEEE Conference on Data Mining.*

[22] H. Wang, A. Klaer, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[23] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural svm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1422–1431, 2013.

[24] X. Yang and Y. Tian, "Action recognition using super sparse coding vector with spatio-temporal awareness," in *ECCV*, 2014, pp. 727–741.

[25] X. Xu, I. Tsang, and D. Xu, "Soft margin multiple kernel learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 749–761, 2013.

[26] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *ECCV*, 2014, pp. 581–595.

[27] J. Liu, J. Luo and M. Shah "Recognizing realistic actions from videos "in the wild"," 2009 IEEE Conference on Computer Vision and Pattern Recognition, DOI: 10.1109/CVPR.2009.5206744.