A Three Stream Deep Network on Extracted Projected Planes for Human Action Recognition

Suraj Prakash Sahoo, Samit Ari
Department of Electronics and Communication Engineering
National Institute of Technology, Rourkela, India
surajprakashsahoo@gmail.com, samit@nitrkl.ac.in

***Abstract*—Human actions are challenging to recognize as it varies its shape from different angle of perception. To tackle this challenge, a multi view camera set up can be arranged, however, it is not cost effective. To handle this issue, a multi stream deep learning network is proposed in this work which is trained on different 3D projected planes. The extracted projected planes which represents different angle of perception, are used as an alternative to multi view action recognition. The projected planes are such that they represents top, side and front view for the action videos. The projected planes are then fed to a three stream deep convolutional neural network. The network uses transfer learning technique to avoid training from scratch. Finally, the scores from three streams are fused to provide the final score to recognize the query video. To evaluate the proposed work, the challenging KTH dataset is used which is widely used and publicly available. The results show that the proposed work performs better compared to state-of-the-art techniques.**

Index Terms*—*Convolutional neural network, projected planes, score fusion, transfer learning.

I. Introduction

The motivation behind human action recognition (HAR) is to extract human behavior from a scene automatically to detect abnormality during surveillance. The example of some of the abnormal situations are illegal entry to restricted area, fighting, accident etc. The present surveillance architecture is totally dependent on human security personnel to analyze the camera output. It needs careful observation and tedious 24×7 human resource availability. If the security personnel becomes sleepy or careless, he/she may skip some abnormal happenings as illustrated in Fig. 1. Thus, it motivates the machine learning and computer vision researchers to develop an automatic HAR paradigm, which can reduce burden from security personnel. The developed HAR can also be applicable in recognizing patient's behavior in health care.

In literature, the techniques for HAR can be divided into two categories according to the feature extraction approaches : HAR through handcrafted feature [1]–[4] or automatic learned deep features [5]–[7]. The work of [1] represents an action video by extracting spatio-temporal interest points (STIPs). HOG [2] and HOF [3] features are used for the holistic representation of an action as well as representation of local video patches. Trajectory based HAR is reported in [4]. A single view is not sufficient to handle the action variabilities as reported by [8]. Therefore, they have proposed a view invariant feature which is insensible to camera view point. G. Yu *et al.* [9] have represented the action videos through STIPs and described the local interest points by 2D HOG and HOF. Recently, deep learning techniques are providing better performance in HAR paradigm. To extract temporal relationships in an action, S. Ji *et al.* [7] have proposed a 3D convolution technique in deep learning. Krizhevsky *et al.* [5] have developed AlexNet architecture for image classification and trained the network with ImageNet dataset.

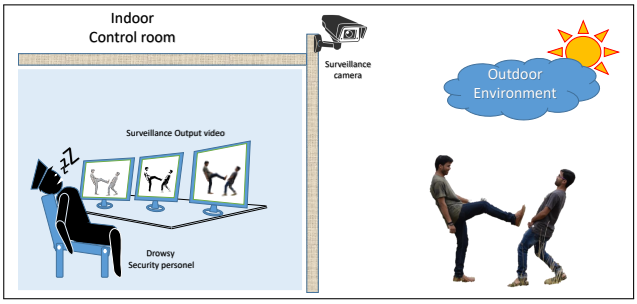

Fig. 1: An illustration of application of HAR in real time environment.

From the literature, it is clear that the reported algorithms have not taken any special attention to distinguish closely related actions. Deep learning networks such as convolutional neural network (CNN) have produced better performance. However, adjusting the kernel weights to distinguish closely related actions is a tedious task. The closely related actions can be viewed differently from different view angle. However, to set up a multi-camera multi-view set up is not cost effective. To overcome this challenge, a HAR algorithm is developed in this work which is based on training of a three stream deep CNN on extracted projected planes representing top, side and front view. Projected planes are extracted from background subtracted action frames by projecting them onto a single plane. Thus, it can be theorized that projected planes are extracted to analyze the action shape from different view points. The projected planes are fed to three pre-trained AlexNet separately to finetune the weights through transfer learning. These three trained network on different projected planes are used during testing for action recognition and score generation. Finally, scores from all the three networks are fused to provide the final recognition score. The work is evaluated on well established KTH dataset and compared with state-of-the-art techniques.

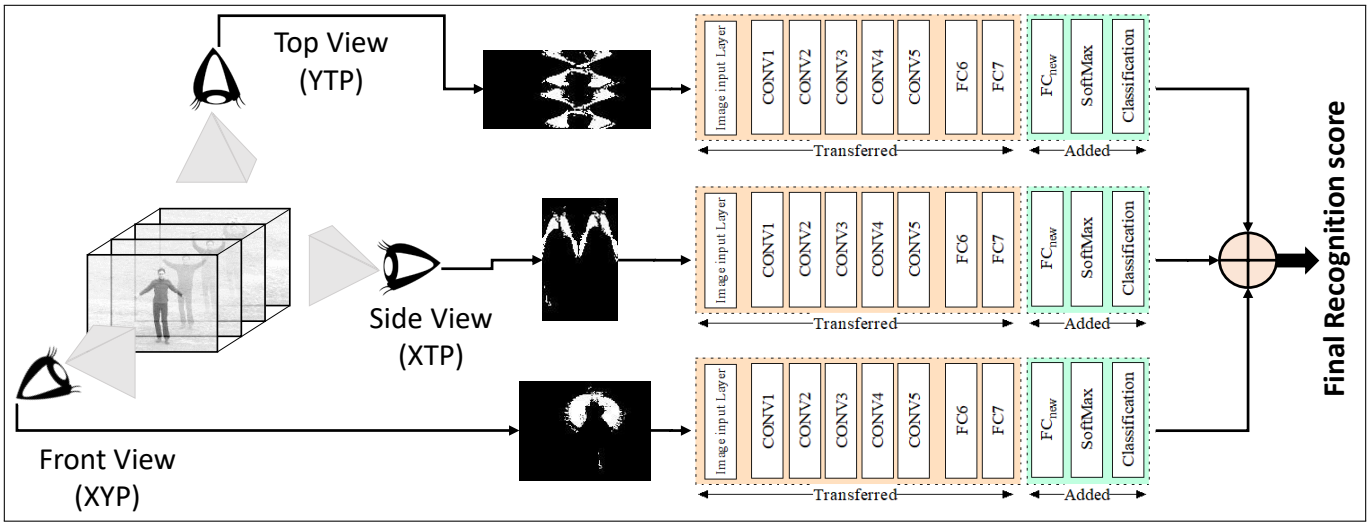

Fig. 2: Block diagram of the proposed three stream deep network for HAR.

Remaining paper is structured in three further sections. The proposed work is explained in section II. The developed algorithm is evaluated on benchmarked dataset and discussed in section III. Finally, the work of this paper is concluded in section IV.

II. Proposed Framework

The block diagram of the proposed three stream deep network is depicted in Fig. 2. The overall procedure is divided into two parts: extraction of different projected planes and training of deep networks on these extracted planes. Three type of projected planes are extracted to represent top, side and front view of the action video. These three projected planes are used to train a three stream network for HAR. Finally, the scores from three stream are fused to provide the final score for action recognition.

A. 3D-projected planes

The projected planes are extracted from the action video which symbolizes different view angles. Action is a 3 dimensional data which can be viewed from top (YT plane), front (XY plane) or side view (XT plane) [10]. The different view planes for an object is illustrated in Fig. 3. The closely related actions such as run, jog, and walk can be described distinguishably by leveraging XT and YT planes as the actions are only separated by speed not by shape. The procedure extracts human silhouette from each frame and then project it to corresponding view planes to describe an action. The advantage of XT and YT projected planes can be explained through the Fig. 4. From the figure it is clearly observed that, the XT plane provides information about the duration of action happening. Similarly, YT plane gives the duration which can be distinguished by the angle of the white patch. For comparison, same number of action frames are considered for all three actions to form XTP and YTP. As running is the fastest among the three actions, its action duration is the

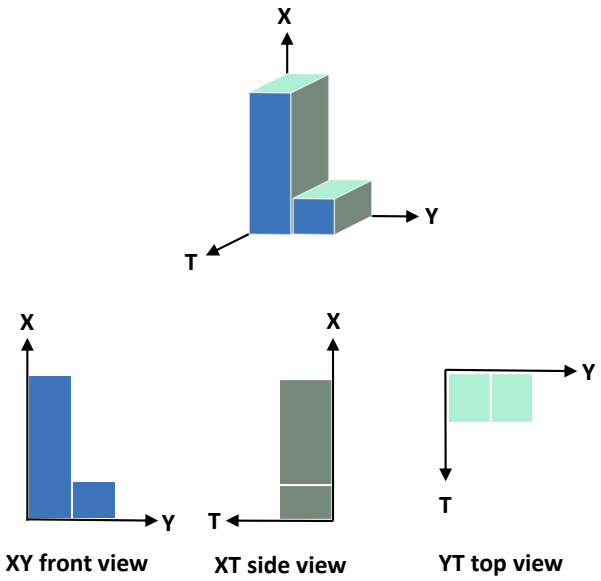

Fig. 3: Illustration of different view planes for an object such as front view, top view and side view projected onto a single plane.

lowest as shown in Fig. 4(d). Similarly, if top view or YTP is analyzed, it can be seen that running action ends faster and thus, its slope of white patch is less compared to other actions. The detailed algorithm for constructing of XTP, XYP and YTP is shown in Algorithm 1.

B. Three stream network with transfer learning

Deep learning architectures are most effective techniques to train action frames or action videos. However, the training needs huge training data to learn the kernel weights effectively. Training on small training data will lead to overfitting of the network. To handle this problem, transfer learning tech-

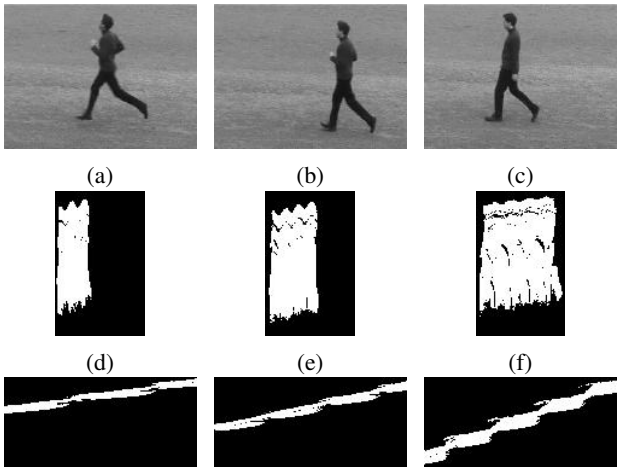

Fig. 4: Extracted XT and YT projected planes alongwith original action frames. (a-c) original action frames of run jog and walk, (b) projected XT planes for run jog and walk, (c) projected YT planes for run jog and walk.

nique [6] can be helpful. The technique leverages the learned knowledge of a pre-trained deep learning network, trained on bigger datasets. In our proposed work, AlexNet is used for transfer learning purpose. The proposed technique uses three pre-trained AlexNet and finetunes their weights through XTP, XYP, YTP. The AlexNet is chosen over other complex and good deep architectures because of its simplicity and low complexity. The AlexNet architecture contains 5 convolution layers and 3 fully connected layers. For transfer learning, the final fully connected layer along with classification and softmax layer are replaced according to the new experiment. The detailed architecture of the AlexNet architecture is shown in Table I.

C. Score fusion technique

The score fusion technique fuses the scores from three different streams to provide the final recognition score. Let, S_{XT}, S_{YT}, S_{XY} be the scores generated from three streams. Then, the final score is generated as follows:

$$S_{final} = \alpha S_{XT} + \beta S_{YT} + \gamma S_{XY} \tag{1}$$

Here, $\alpha + \beta + \gamma = 1$. The values are adjusted in such a way that it provides better result.

III. Results and Discussion

The proposed technique is evaluated on KTH dataset [11] which contains similar actions like 'run', 'jog', and 'walk'. The other actions of the dataset are 'punch', 'clap', and 'hand waving'. The total number of action videos in KTH dataset are 598 as each class is having 99-100 videos. Each action class comprises of actions from 25 different persons in indoor and outdoor environments. During training, actions of 16 persons are fed to the proposed architecture. The testing dataset is comprising of the actions of rest 9 persons. All the experimentation work is carried out in MATLAB 2018a platform with windows 10 operating system. The implementations are boosted by the use of NVIDIA Quadro M4000 8GB GPU card.

Algorithm 1: Extraction of 3D projected planes

Initialize: *V(X,Y,T)* as the human action video;

Procedure $Frames\_binaralization(V(X,Y,T))$
while $i < T$ **do**
  $D(i) = MedianFilter(V(:,\ :,\ i+1) - V(:,\ :,\ i))$;
  $i = i$+1;
end

Procedure $XT\_ProjectedPlane(D(X,Y,T))$
XTP=zeros(X,T);
for *l=1:T* **do**
  Step 1 : Select the action frame *D*(: , : , *l*)
  Step 2 : Add pixels of all columns in a row
  Step 3 : Save it to XTP(: , *l*)
end

Procedure $YT\_ProjectedPlane(D(X,Y,T))$
XTP=zeros(T,Y);
for *l=1:T* **do**
  Step 1 : Select the action frame *D*(: , : , *l*)
  Step 2 : Add pixels of all rowss in a column
  Step 3 : Save it to YTP(: , *l*)
end

Procedure $XY\_ProjectedPlane(D(X,Y,T))$
XYP=zeros(X,Y);
for *l=1:T* **do**
  Step 1 : Select the action frame *D*(: , : , *l*)
  Step 2 : Add it to XYP
end
Step 3 : Normalize XYP

The discussion starts with the confusion matrix of the KTH dataset provided by the proposed three stream technique. The diagonal cells contains the accuracy of the each action class *i.e.* true positive (TP). The column cells except diagonal cells are false positive (FP) values. Similarly, the row cells represents the false negative (FN) values. The accuracy is calculated by dividing total number of testing action videos to correctly recognized action videos. The overall accuracy is found to be 92.01% for the proposed algorithm. From confusion matrix, it is clear that the *'walking'* action is better classified. However, some of the closely related actions (run, jog, walk) are still being confused with each other. As a result the overall accuracy is decreasing.

The better recognized action class can not be decided only by accuracy parameter. Other parameters like sensitivity (Sen), specificity (Spe), and precision (Pr) are calculated for each class as follows:

$$Sen = \frac{TP}{TP + FN} \times 100 \tag{2}$$

TABLE I: AlexNet architecture with layer details used for transfer learning.

Sl No.	Layers	Details	Filter size	Stride	Number of Channels
1	Convolution Layer 1	Conv1	11×11	[4, 4]	96
		Max Pool1	3×3	[2, 2]	
2	Convolution Layer 2	Conv2	5×5	[1, 1]	256
		Max Pool2	3×3	[2, 2]	
3	Convolution Layer 3	Conv3	3×3	[1, 1]	384
4	Convolution Layer 4	Conv4	3×3	[1, 1]	384
5	Convolution Layer 5	Conv5	3×3	[1, 1]	256
		Max Pool5	3×3	[2, 2]	
6	Fully connected	fc6			4096
7	Fully connected	fc7			4096
8	Fully connected	fc8			1000

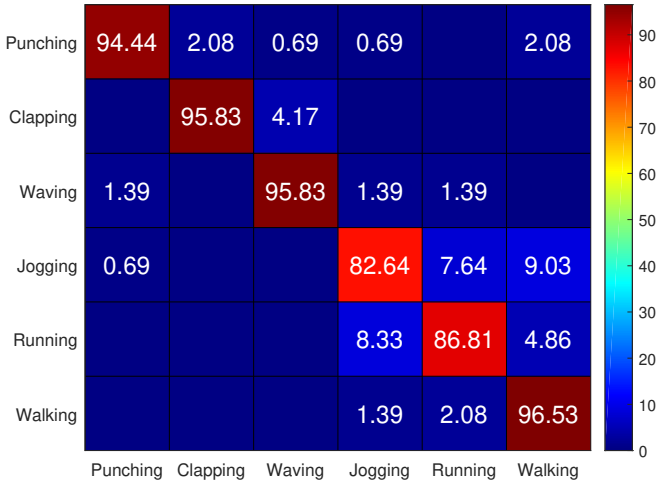

Fig. 5: Confusion matrix for the KTH dataset for the proposed technique.

$$Spe = \frac{TN}{TN + FP} \times 100 \tag{3}$$

$$Pr = \frac{TP}{TP + FP} \times 100 \tag{4}$$

The calculated parameters for each action class are shown as a bar diagram in Fig. 6. For *'Clapping'* action, the specificity and precision values are 0.9958 and 0.9788 respectively which are better compared to other classes. The sensitivity value is also comparable to other classes which concludes that the *'Clapping'* action is the better recognized class.

The result of the proposed method is compared to state-of-the-techniques [8], [9], [12], [13] in Table II. K.P. Chou *et al.* [8] have worked towards view invariant features when multi view human action are there. G. Yu *et al.* [9] have used tree structures to index the local features. They have used local 2D HOG and HOF features in their work. 2D features and view invariant features are not having any special approach to distinguish the actions differentiated by action time. The reported techniques in [8], [9] have not taken any special

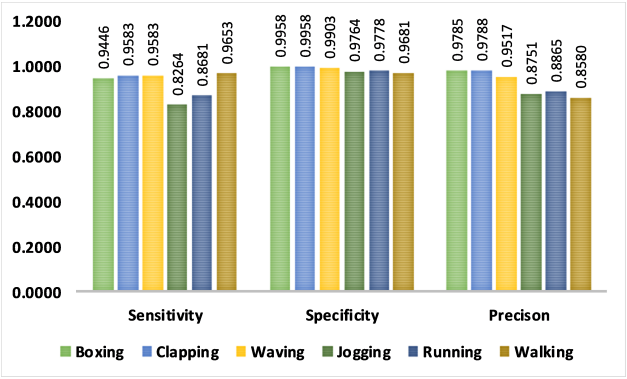

Fig. 6: Classwise statistical indices for each action class of KTH dataset.

care to handle the closely related actions. In this work, view based projected planes are extracted and used to train separate AlexNet networks to finetune their weights. The effect of combinations of XT, XY and YT planes are studied in this proposed work. An overall accuracy of 92.01% is achieved from the proposed method which is better compared to the state-of-the-art techniques.

TABLE II: Comparison of action classification on KTH dataset with state-of-the-art methods using proposed technique

Method	**Accuracy** (%)
Ikizler-Cinbis *et al.* [12]	81.17
Wu *et al.* [13]	83.30
K P Chou *et al.* [8]	90.58
Yu *et al.* [9]	91.80
Proposed method	**92.01**

IV. Conclusions

In this paper a three stream CNN is proposed which is trained on 3D projected planes. The 3D projected planes are used as the alternative to multi-view action recognition. Three types of projected planes are extracted such as top view or YTP, side view or XTP and front view or XYP. The transfer learning technique on pre-trained AlexNet network is carried

out to fine tune its weights. The trained network is then used to recognize human actions. Score fusion technique is used to fuse three scores from three stream to provide the final recognition score. The work is evaluated on publicly available KTH dataset. An performance accuracy of 92.01% is achieved with the proposed technique which is better compared to state-of-the-art techniques.

ACKNOWLEDGMENT

The work is supported by Digital India Corporation (formerly Media Lab Asia) under Visvesvaraya Ph.D. Scheme for Electronics and IT under the department of MeitY government of India. [grant number PhD-MLA/4(13)/2015-16]

REFERENCES

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1, 2005, pp. 886–893.

[3] J. L. Barron and N. A. Thacker, "Tutorial: Computing 2D and 3D optical flow," *Imaging Science and Biomedical Engineering Division, Medical School, University of Manchester*, vol. 1, 2005.

[4] Y. Yuan, Y. Zhao, and Q. Wang, "Action recognition using spatial-optical data organization and sequential learning framework," *Neurocomputing*, vol. 315, pp. 221–233, 2018.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[6] D. Han, Q. Liu, and W. Fan, "A new image classification method using cnn transfer learning and web data augmentation," *Expert Systems with Applications*, vol. 95, pp. 43–56, 2018.

[7] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[8] K.-P. Chou, M. Prasad, D. Wu, N. Sharma, D.-L. Li, Y.-F. Lin, M. Blumenstein, W.-C. Lin, and C.-T. Lin, "Robust feature-based automated multi-view human action recognition system," *IEEE Access*, vol. 6, pp. 15 283–15 296, 2018.

[9] G. Yu, N. A. Goussies, J. Yuan, and Z. Liu, "Fast action detection via discriminative random forest voting and top-k subvolume search," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 507–517, 2011.

[10] R. Silambarasi, S. P. Sahoo, and S. Ari, "3D spatial-temporal view based motion tracing in human action recognition," in *IEEE International Conference on Communication and Signal Processing (ICCSP)*, 2017, pp. 1833–1837.

[11] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *IEEE International Conference on Pattern Recognition (ICPR)*, vol. 3, 2004, pp. 32–36.

[12] N. Ikizler-Cinbis and S. Sclaroff, "Web-based classifiers for human action recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1031–1045, 2012.

[13] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 236–243, 2013.