Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-11220-4

1220: VISUAL AND SENSORY DATA PROCESSING FOR REAL TIME INTELLIGENT SURVEILLANCE SYSTEM

Deep convolutional neural model for human activities recognition in a sequence of video by combining multiple CNN streams

Neeraj Varshney[1] · **Brijesh Bakariya**[2]

Received: 9 March 2021 / Revised: 10 June 2021 / Accepted: 6 July 2021

Abstract
The video file is a collection of image sequential; this image sequence holds both spatial and temporal information. Optical flow and motion history images are two well-known methods for the identification of human activities. Optical flow describes the speed of every individual pixel point in the picture. Still, this information about the motion cannot represent the complete action and different movement speeds. The durations of Local body parts show almost similar intensity in the Motion history image. Therefore, similar actions are not identifying with good precision. In this paper, a deep convolutional neural model for human activities recognition video has been proposed in which multiple CNN streams are combined. The model combines spatial and temporal information. Two fusion schemes, i.e. Average fusion and convolution fusion of spatial and temporal stream, are discussed in this paper. The proposed method performs better than other approaches based on human activity recognition methods on a benchmark dataset, namely UCF101 and HMDB51.Average fusion score 95.4% test accuracy and convolution fusion score 97.2% test accuracy on UCF101 and for HMDB51, average fusion score 84.3% and convolution fusion score 85.1% respectively.

Keywords Human activity recognition · Features extraction · Deep convolution neural network · Spatial and temporal · Convolution fusion · Average fusion

✉ Neeraj Varshney
neeraj.varshney@gla.ac.in

Brijesh Bakariya
dr.brijeshbakariya@ptu.ac.in

[1] I. K. Gujral Punjab Technical University, Kapurthala, India

[2] I. K. Gujral Punjab Technical University, Hoshiarpur Campus, Hoshiarpur, India

Published online: 26 August 2021

1 Introduction

The limited human capabilities to analyze videos in a natural way demands for the intelligent systems that can automatically analyze and recognize activities or events occurring in videos like surveillance, health care, monitoring the old people at home, robotic vision, crowd monitoring etc. Scene identification is one of the crucial tasks in a video. Numerous CNN-based state-of-the-art methods [2, 6, 7–9, 14, 17, 19, 27] are available for activity recognition but suffer from challenges to provide the desired result. The accuracy of the activity classification using the Convolution network can be additionally upgraded by fusing different streams [26]. Some activity can be identified by spatial information (via appearance) and some required information about the motion [13]. The two most common methods to detect motion are Motion history image (MHI) and Optical Flow (OF). The durations of Local body parts show the same intensity in the MHI. Therefore, it is difficult to identify similar actions. Whereas OF describes the motion between two consecutive video frames, it is insufficient to describe the complete action. Therefore, combining spatial and temporal information needs an hour to effectively identify activity and handle the issues that occur during the collection of motion information via state-of-the-art techniques like OF and MHI. Here, the Proposed work keeps attention on combining the spatial and temporal stream of CNN efficiently so that it can be feasible computationally and may be used in real-time applications. This paper proposes an approach based on a convolution neural network, extracting spatial and temporal information and fusing it to CNN. Two different kinds of Fusion scheme are proposed. Namely, average fusion and convolution. The detailed procedure of fusion is discussed in Sect. 5.

2 Related work

The task of action/ activity recognition involves the identification of activity from video which perform action for certain duration in the video. In contrast to the image processing the video processing required the spatiotemporal information as well with spatial information.

Karpathy et al. [8] proposed multiple ways for fusing the temporal information using 2D convolution network. Author used several fusion scheme in their experiment. Single architecture for single frame that fuses information collected from all frames at the last stage of the ConvNet, late fusion, early fusion, slow fusion but the accuracy result was not up to the mark because the motion features were missing in the spatiotemporal features. Simonyan and Andrew [14] proposed the availability of motion feature while designing the model. They used two separate network in their architecture one for spatial domain and another for spatiotemporal. the author use frames for the input to the spatial network and for the temporal network optical flow stacked for 10 continuous frames are used. They fuse both spatial and temporal network at the last convolution layer, method improve the issues of single stream. the long range temporal information was missing in learnt features because predictions were obtained from averaging predictions over sampled clips also the method involves pre computation of optical flow vector and store separately which cause extra computation. Tran et al. [19] proposed an approach for spatiotemporal feature learning using deep 3-dimensional convolutional networks as 3D ConvNets perform better for spatiotemporal feature compared to 2D ConvNets. Sun et al. [17] proposed concept of factorized 3D ConvNet, author proposed an idea to break 3D convolutions into spatial stream 2D convolutions followed by temporal stream 1D convolutions. The 1D ConvNet placed just after 2D layer, was implemented as 2D

convolution over temporal and channel dimension. Feichtenhofer et al. [5] propose fusion of spatial and temporal network at convolution layer. Combining temporal net output across time frames so that long term dependency is also modelled. Girdhar et al. [7] proposed a learnable feature aggregation (VLAD) instead of aggregation using max pool or avg pool. Zhu et al. [27] majorly focus on improving accuracy and associated cost of majoring activity. The authors explored multiple strategies and architectures to generate optical flow with largest fps and least parameters without hurting accuracy much. They discuss the usage of an unsupervised architecture to generate optical flow for a stack of frames. Khurana and Kushwaha [9] used fusion-based dual-stream deep model for HAR. Author apply 2D convolution for spatial data and 3D for the spatiotemporal and decision level fusion done at the last FC layer of the CNN. Khaire et al. [10] Fuse the softmax score of Separately train MHI, DMM and skeleton images at the classification level for activities recognition using RGB-D dataset. A view invariant silhouette-based recognition of human activity was proposed by Kushwaha et al. [12]. Singh et al. [15] also proposed to identify multi-view activity using HMM. Tu et al. [20] proposed 3 different model based on two-stream using motion and appearance. Wang et al. [23] proposed three-stream model, in which they consider spatial stream, global features and local features for the model.

3 System overview

Our literature study recognises many multi-stream neural networks with variable fusion schemes for activity recognition but identifies many challenges. The training cost of deep learning is the critical factor behind the motivation of our work. So, we use network weights with ImageNet [3], as it is rare to have a dataset of enough size to train the network with random initialization of weights. The easiest methodology for applying CNN for action recognition from the recorded video is to treat video as still pictures as video streams comprise a grouping of video stream across temporal measurement. Individual frame-level activity can be easily identified by the convolution neural network but this methodology does not contain the information about the motion sequence of the frames, so motion information is just overlooking during the activity recognition. Some activities can be conceivable from appearance just, but some required temporal information for identification. Temporal information can be identified in many ways like motion energy image, MHI, Optical flow but all have their limitation. Moving towards the direction of multi-stream models, we implement a multi-stream (Spatial + Temporal) model for activity recognition. Our proposed approach calculates temporal information, which is extracted using optical flow-motion history image [OF-MHI] proposed in [18] and provide to the CNN model.

The architecture model of the multi-stream model comprises a fusion of spatial and temporal stream. Figures 1 and 2 shows an overview of our work. Figure 1 describes our proposed method in which there are two streams, spatial and temporal which are later fused using the average fusion method as presented in Algorithm 1. Whereas Fig. 2 describes our method in which two streams are fused using the convolution fusion. Weights are initialized with pre-trained models from ImageNet. Based on the convolution fusion and average fusion of spatial and temporal stream, the final class of the activity is calculated. Motion is significant to find the action. Once we detect and isolate an object or person of interest, we can extract valuable data such as positions, velocity, acceleration, and so on. This information can be used for action recognition, behaviour pattern studies, video stabilization, augmented reality, and so on.

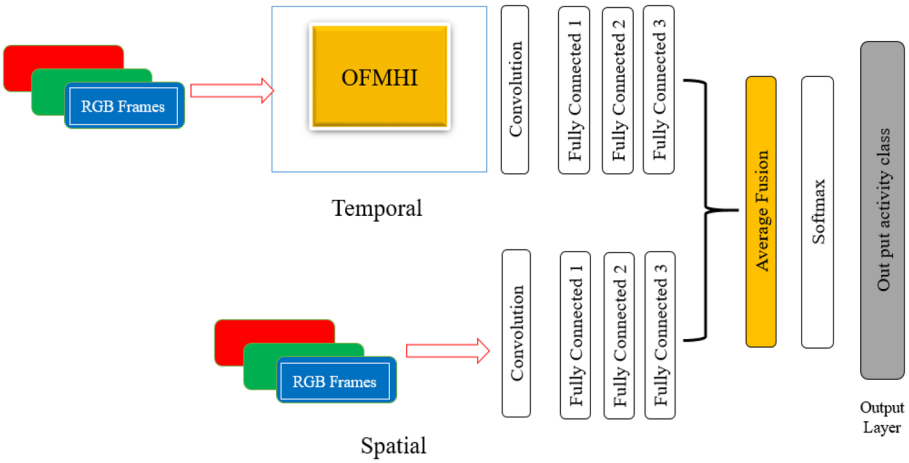

Fig. 1 Proposed model with Avg. fusion

The Optical Flow technique is a pattern of an object's apparent motion. Surfaces and edges in a visual scene are caused by relative motion between an observer and scene or between the camera and the scene. The speed of every individual pixel point in the picture can easily describe by optical flow. But, this information about the speed is not able to represent the complete action. The motion history image (MHI) is a static layout to describe the movement area and path as a single image template. Different movement speeds and durations of Local body parts show the same intensity in the MHI. Therefore, identification of similar action become tough and indistinguishable. Therefore, the motion quality of every pixel point is gathered by the optical stream length at that area and then exponentially rationalized over time. The intensity value of each foreground pixel is fixed in MHI, no matter about the time and speed of the pixel so local movement can easily describe with it but very difficult to identify the similar action like walking or running, sitting and getting up [18]. Video contains both spatial and temporal components, in which the spatial part carries information about the scene and the temporal part contains information about the movement of the object in the scene.

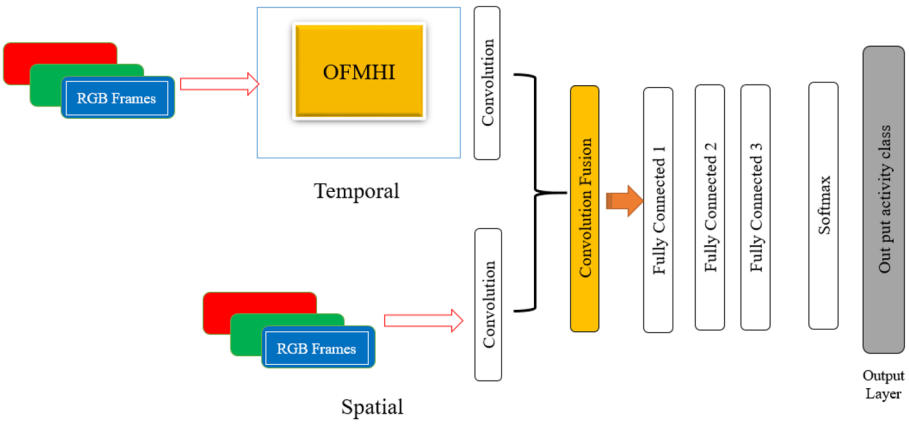

Fig. 2 Proposed model with Convolution fusion

4 The proposed solution comprises of these two network streams

"Spatial stream" and "temporal stream". In spatial stream RGB frames are provided and in temporal stream OF-MHI (Optical Flow-Motion history image) is provided which gives the motion information of the activity. "Spatial stream" provide activity representation from RGB frames whereas temporal stream provides the information about the motion in the activity. We have fused these two streams using average fusion and at convolution layer just before fully connected layer to predict the activity as proposed in Sects. 4.1 and 4.2.

4.1 Average fusion method

In Average Fusion method both spatial streams and temporal streams are fused to get the average of both the classes. We apply three FC layer in both the stream. Detailed layers and dimensions about the layers are discussed in Sect. 4.1.1

4.1.1 Train/test for spatial stream

It works on individual frames of the video frames and perform action recognition effectively these individual frames. A sub-image of 224×224 is cropped from each frame. certain activities are strongly connected with object appearance therefore static appearance itself is a useful indication to identify an activity. We use the batch size of 200 and for learning the network parameter, stochastic gradient descent algorithm is use and momentum set to 0.8, also weights are initializing with pre-trained models from ImageNet [3]. We set the learning rate 0.001.

4.1.2 Train/test settings for temporal stream

For temporal stream we are using modified version of MHI proposed in [18]. A fixed intensity value t is assign to each foreground pixel in the traditional MHI which shows the same pattern in the fast and slow movement of different body parts. Therefore with each pixel value (x,y) is associate with the optical flow length s(x, y) over time.

The temporal feature is calculated by

$$\mathbf{E}(\mathbf{x}, \mathbf{y}, \mathbf{t}) = \mathbf{s}(\mathbf{x}, \mathbf{y}, \mathbf{t}) + \mathbf{E}(\mathbf{x}, \mathbf{y}, \mathbf{t} - 1) \cdot \boldsymbol{\alpha} \tag{1}$$

where.

S (x, y, t) = optical flow length of pixel at time t.

$\boldsymbol{\alpha}$ = Update rate (Range 0 to 1).

4.1.3 Convolution

Two Convolution layers are added having dimensions of 64 channel with kernel size of 3×3 and same padding. Pooling (Maxpool) of 2×2 and same stride is used. Two layers of 128 channel with 3×3 kernel and same padding, pooling of 2×2 and same stride used here also. Likewise, three layers of 256 channel with 3×3 kernel and same padding and 2×2 pooling and stride 2×2 is used. Three layer of 512 channel with 3×3 kernel and same padding, with 2×2 pooling and same stride is used. 3 FC layers are used before fusion.

4.1.4 Average fusion

Our proposed average fusion model is inspired by the VGG model and referred from [16]. The block architecture of proposed spatial and temporal Convolution network is shown in Fig. 1. Convolution layers of 3×3 filter used in the model with a stride 1 and same padding and max-pool layer of 2×2 filter of stride 2 is use. Such organization of convolution and max pool layers follow in the whole network. We use the same architecture for both spatial stream as well as temporal stream. Then we take the average score of both the class at the end for the classification of the activity.

ALGORITHM 1: AVERAGE FUSION

Input: A training dataset $Dtrain = \{X^{trg}, Y^{trg}\}$, and a testing dataset $Dtest = \{X^{ted}\}$
(input size: 224 x 224 x 3;)
Output: Activity labels Y^{ted} of the unlabeled testing data
/* pre-processing */
1 Segment both the training and test data into sequences to form two sets of input vectors $X^{trg} = \{x_1, \cdots, x_{N^{trg}}\}$, and $X^{ted} = \{x_1, \cdots, x_{N^{ted}}\}$;
2 Standardize and whiten the training input vectors and then applying on the test input vectors;
3 Make batches;
4 Convolution
    /* **Spatial Domain** */
    i. Conduct the convolution operation on the input data;
    ii. Apply the ReLU function on the convolutional layer output;
    iii. Conduct the max-pooling operation on the ReLU layer output;
    iv. Flatten the pooled results into a vector and feed it to the FC layer1;
    v. Fully connected Layer 2;
    vi. Fully connected layer 3;
    /* **Temporal Domain** */
    i. Conduct the convolution operation on the input data;
    ii. Apply the ReLU function on the convolutional layer output;
    iii. Conduct the max-pooling operation on the ReLU layer output;
    iv. Flatten the pooled results into a vector and feed it to the FC layer1;
    v. Fully connected Layer (FC) 2;
    vi. Fully connected layer (FC) 3;
5. Apply Average Fusion
/* Classification */
6. Use the trained network to predict the labels Y^{ted} of test data X^{tet}

4.2 Convolution fusion method

In Convolution Fusion method both spatial streams and temporal streams are fused at the last convolution layer and after fusion apply three FC layer and then apply SoftMax layer for activity recognition.

4.2.1 Spatial stream

It operates on individual video frames same as in the average fusion, effectively performing action recognition from still images. Same as used in average fusion.

4.2.2 Temporal stream

For temporal stream we are using modified version of MHI proposed in [18]. Same as used in average fusion.

4.2.3 Convolution

We added two Convolution layers are added having dimensions of 64 channel with kernel size of 3×3 and same padding. Pooling (Maxpool) of 2×2 and same stride is used. Two layers of 128 channel with 3×3 kernel and same padding, pooling of 2×2 and same stride used here also. Likewise, three layers of 256 channel with 3×3 kernel and same padding and 2×2 pooling and stride 2×2 is used. Three layer of 512 channel with 3×3 kernel and same padding, with 2×2 pooling and same stride is used. 3 FC layers are used before fusion.

4.2.4 Convolution fusion

The block architecture of Convolution fusion is schematically shown in Fig. 2. Model consist convolution layers always used same padding and maxpool layer of 2×2 and stride 2. This arrangement is followed consistently throughout the whole architecture. We fuse both the class at the last layer of convolution and use 3 FC layers (Fully Connected) and then softmax layer and classify the activity.

ALGORITHM 2: CONVOLUTION FUSION

Input: A training dataset $Dtrain = \{X^{trg}, Y^{trg}\}$, and a testing dataset $Dtest = \{X^{ted}\}$
(input size: 224 x 224 x 3;)
Output: Activity labels Y^{ted} of the unlabeled testing data
/* pre-processing */
1 Segment both the training and test data into sequences to form two sets of input vectors $X^{trg} = \{x_1, \cdots, x_N{}^{trg}\}$, and $X^{ted} = \{x_1, \cdots, x_N{}^{ted}\}$;
2 Standardize and whiten the training input vectors and then applying on the test input vectors;
3 Make batches;
4 Convolution
/* **Spatial Domain** */
i. Conduct the convolution operation on the input data;
ii. Apply the ReLU function on the output of the convolutional layer;
iii. Conduct the max-pooling operation on the output of the ReLU layer;
/* **Temporal Domain** */
i. Conduct the convolution operation on the input data;
ii. Apply the ReLU function on the output of the convolutional layer;
iii. Conduct the max-pooling operation on the output of the ReLU layer;
5. Apply Convolution Fusion (at the last convolution layer)
i. Flatten the pooled results into a vector and feed it to the FC layer1;
ii. Fully connected Layer 2;
iii. Fully connected layer 3;
6. Apply Softmax layer for activity recognition
/* Classification */
7. Use the trained network to predict the labels Y^{ted} of test data X^{tet}

Fig. 3 Few snapshots example of RGB frames from UCF101 [16]

5 Dataset and performance measure

In this paper two widely used datasets UCF101 and HMDB51 are for the experiment and result analysis. These two datasets are widely used by many states of art papers for their experiment.

5.1 UCF101 dataset

UCF101 dataset [16] is widely used by the researcher in the field of human activity recognition. There are 101 action classes in it with total 13,320 videos. Some of the activity class videos shown in Fig. 3. UCF101 is a diversify dataset with variation in background, camera motion, scale, viewpoint, pose, lighting condition etc., These diversify condition make research more challenging and effective. The recordings in 101 activity classifications are gathered into 25 groups, there are 4–7 videos of an action, in each group. Similar group videos contain similar features like viewpoint, background etc. The action categories include five types:

1. human interaction with object,
2. human body movement,
3. sports activity
4. musical instruments play and
5. interaction between human–human,

Fig. 4 Few snapshots example of RGB frames from HMDB51 [11]

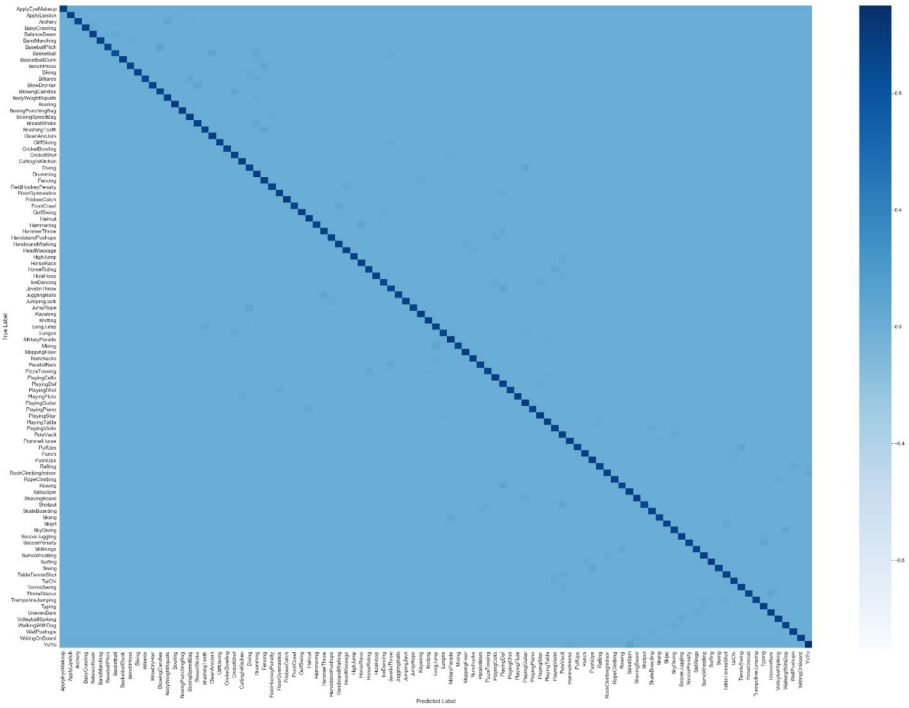

Fig. 5 Confusion matrix of multi stream convolution fusion on UCF101

5.2 HMDB51 dataset

Another dataset we examine is HMDB51 [11]. Videos in HMDB51 are collected from multiple sources, typically from movies clips, clips from the database like prelinger archive, YouTube and Google videos are also included, Fig. 4. HMDB51 contains total 51 action categories with at least 101 clips each. Total 6849 video clips are there in the dataset. Dataset consist five major action groups General Human facial actions (2) Facial actions with object handling (3) Human body movements (4) Body movements with object interaction (5) Body movements for interaction with human.

6 Experiment result and discussion

In this section, we discuss the result of proposed model. We use two datasets to compared with the-state-of-art methods for human activity recognition: UCF101 dataset and HMDB51dataset. Proposed model is trained and tested on these two datasets for activity recognition.

UCF101 [16] is very commonly used video dataset for activity recognition. It consists of 101 classes of various activities with total 13,320 videos. There are 25 groups of videos consist of 4–7 videos of an action. Table 1 describe the detail of UCF101.

Another dataset is HMDB51 [11], This dataset consists videos from various sources. These videos are from movies clips, and collection from some open databases Table 2 describe details of HMDB51 dataset.

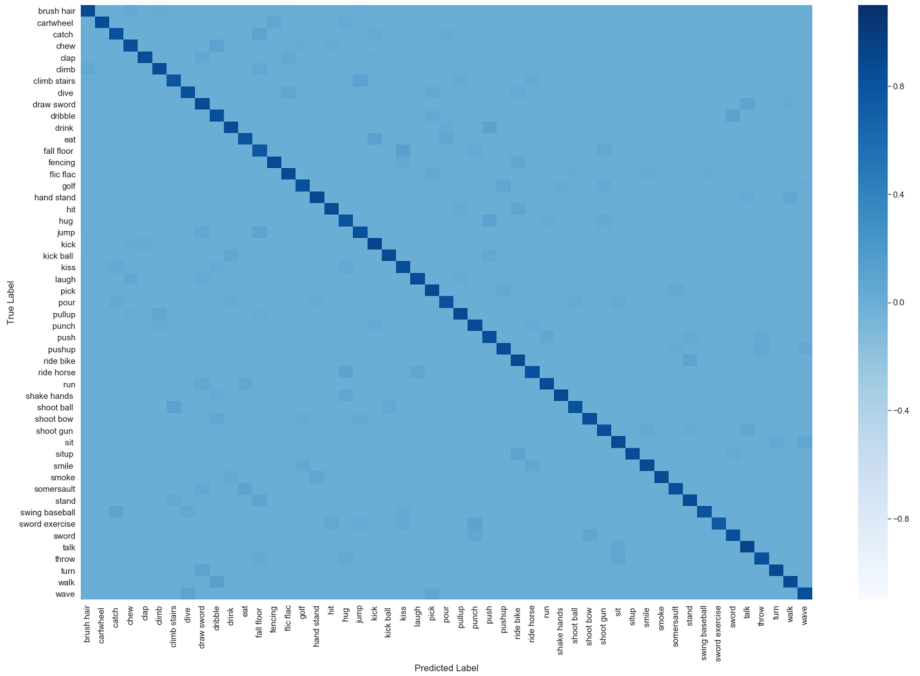

Fig. 6 Confusion matrix of multi stream convolution fusion on HMDB51

6.1 Performance of various state of art method on UCF101 dataset

Table 3 shown the efficiency of the average fusion and convolution fusion used in proposed model on UCF101 dataset also various state of art approaches compare with recognition results on the UCF101 dataset in Table 4. From this table, we find that accuracy is improved using our proposed model, and believed that it can be further improved by using more fusion strategy. Figures 5 and 6 shows the confusion matrix of the convolution fusion model for UCF101 and HMDB51 dataset respectively.

Classification accuracy on UCF dataset is shown in Table 4 and compared with techniques based on handcrafted feature methods, early fusion, late fusion and deep convolution neural network. Both fusion method outperforms. The convolution fusion is ~2% better as compared to [25] and ~7% better in comparison to [4].

Table 1 Summary of UCF 101 dataset

Action Class	101
Total video	13,320
Per action groups	25
Total time	1600 minuets
Max clip size	71.04 Sec
Min clip size	1.06 Sec
Avg clip size	7.21 Sec
Resolution	320×240
Year	2012

Table 2 Summary of HMDB51 Dataset

Action Class	51
Total video	6776
Per action groups	101 at least
Min clip size	1 Sec
Resolution	320×240
Year	2011

Table 3 Result on UCF 101 (split 01)

Method	UCF101 (Accuracy %)
Multi stream (Average fusion)	95.4
Multi stream (Convolution Fusion)	97.2

Table 4 Comparison of our multi-stream model with state-of-the-art methods for UCF101 dataset

Method	UCF101 (Accuracy %)
Slow fusion [8]	65.4
Fusion by Avg [14]	86.2
Fusion by SVM [14]	87.0
Late fusion using VGG m (2048) [5]	85.94
Late fusion using VGG 16 [5]	90.62
Deep dual stream [9]	90
LCRN [25]	82.92
TSN (RGB + Flow)	94.0
TSN (RGB + Flow + Warped flow)	94.2
iDT + FV [21]	85.9
Decision fusion [1]	90.0
Our (Convolution fusion)	97.2
Our (Average fusion)	95.4

Table 5 Result on HMDB51 Dataset

Method	HMDB51 (Accuracy %)
Average fusion	84.3
Convolution fusion	85.1

Table 6 Comparison of our multi-stream model with state of the art methods for HMDB51 dataset

Method	HMDB51 (Accuracy %)
TDD + iDT [24]	65.9
TDD [24]	63.2
Fusion by Avg [14]	59.4
Fusion by SVM [14]	58.0
Two-Stream SVMP(I3D +) [22]	81.3
Two-Stream SVMP(ResNet + iDT) [22]	72.6
Two-Stream SVMP(ResNet) [22]	71
iDT + FV [21]	57.2
Our (Convolution fusion)	85.1
Our (Average fusion)	84.3

6.2 Performance of various state of art method on HMDB51 dataset

Here Table 5 shown the efficiency of the average fusion and convolution fusion on HMDB51 dataset also compare the recognition results to the state-of the art approaches on the HMDB51 dataset in Table 6. Once again the result observe are improved using convolution fusion and average fusion of the spatial and temporal stream. Table 6 consider some of RGB-based and multi-streams methods on HMDB51 for comparison. Both fusion method performs well. The convolution fusion is ~4% better as compared to [22] and ~3% better in comparison to [22].

7 Conclusion

This article presents the recognition of human activity by combining multiple streams. In this work, a pre-trained network on ImageNet uses a spatial stream, and an improved version of MHI for the temporal stream are fused at the last convolution layer and averaging both streams. The proposed model tested on the most promising datasets used in activity recognition, such as UCF101 and HMDB51 and achieved remarkable results. Also, compared proposed model with the other works and found batter result. Results also specified that the proposed model could distinguish similar action with different velocity. Some additional input modality can be used to input the proposed multi-stream networks for activity recognition in future work. The proposed models' performance can be analysed on other benchmark datasets with some more complex activity and can be analysed of multi-view or multiple human activity recognition.

References

1. Bhagat C, Kushwaha AKR (2019) Delving Deeper with Dual-Stream CNN for Activity Recognition: Select Proceedings of IC3E 2018. https://doi.org/10.1007/978-981-13-2685-1_32
2. Bilen H, Fernando B, Gavves E, Vedaldi A, Gould S (2016) Dynamic image networks for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3034–3042

3. Deng J, Dong W, Socher R, Li L, Li K, Li F (2009) ImageNet: a large-scale hierarchical image database. In: CVPR, pp 248–255
4. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
5. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1933–1941
6. Feichtenhofer C, Pinz A, Wildes R (2016) Spatiotemporal residual networks for video action recognition. In: Proceedings of the Advances in Neural Information pro- cessing systems, pp 3468–3476
7. Girdhar R, Deva R, Abhinav G, Josef S, Bryan R (2017) Actionvlad: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 971–980
8. Karpathy A, George T, Sanketh S, Thomas L, Rahul S, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 1725–1732
9. Khurana R, Kushwaha AKS (2019) Delving Deeper with Dual-Stream CNN for Activity Recognition. In Recent Trends in Communication, Computing, and Electronics, pp 333–342. Springer, Singapore
10. Khaire P, Kumar P, Imran J (2018) Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recogn Lett. https://doi.org/10.1016/j.patrec.2018.04.035
11. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: A large video database for human motion recognition. ICCV
12. Kushwaha AKS, Srivastava S, Srivastava R (2017) Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns. Multimedia Syst 23(4):451–467
13. Roy D, Srinivas M, Chalavadi KM (2016) Sparsity-inducing dictionaries for effective action classification. Pattern Recogn. https://doi.org/10.1016/j.patcog.2016.03.011
14. Simonyan K, Andrew Z (2014) Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems, pp 568–576
15. Singh R, Kushwaha AKS, Srivastava R (2019) Multi-view recognition system for human activity based on multiple features for video surveillance system. Multimedia Tools Appl 78(12):17165–17196
16. Soomro K, Zamir AR, Shah M (2012) UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint https://arXiv:1212.0402
17. Sun L, Kui J, Dit-Yan Y, Bertram ES (2015) Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE international conference on computer vision, pp 4597–4605
18. Tsai D-M, Chiu W-Y, Lee M-H (2015) Optical flow-motion history image (OF-MHI) for action recognition. SIViP 9(8):1897–1906. https://github.com/tomar840/two-stream-fusion-for-action-recognition-in-videos
19. Tran D, Lubomir B, Rob F, Lorenzo T, Manohar P (2015) Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pp 4489–4497
20. Tu Z, Xie W, Qin Q, Poppe R, Veltkamp R, Li B, Yuan J (2018) Multi-stream CNN: learning representations based on human related regions for action recognition. Pattern Recogn 79:32–43
21. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3551–3558
22. Wang J, Cherian A, Porikli F, Gould S (2018) Video representation learning using discriminative pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 pp 1149–1158
23. Wang L, Ge L, Li R, Fang Y (2017) Three-stream CNNs for action recognition. Pattern Recogn Lett 92:33–40
24. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-con- volutional descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4305–4314
25. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision, pp 20–36. Springer, Cham. https://towardsdatascience.com/gentle-dive-into-math-behind-convolutional-neural-networks-9a07dd44cf9
26. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: towards good practices for deep action recognition. In: Proceedings of the European Conference on Computer Vision, Springer, pp 20–36
27. Zhu Y, Zhenzhong L, Shawn N, Alexander H (2018) Hidden two-stream convolutional networks for action recognition. Asian Conference on Computer Vision. Springer, Cham, pp 363–378