International Journal of Computer Vision and Image Processing
Volume 12 • Issue 1

Spatio-Temporal Deep Feature Fusion for Human Action Recognition

Indhumathi C., Manonmaniam Sundaranar University, India
Murugan V., Manonmaniam Sundaranar University, India
Muthulakshmi G., Manonmaniam Sundaranar University, India

ABSTRACT

Action recognition plays a vital role in many secure applications. The objective of this paper is to identify actions more accurately. This paper focuses on the two-stream network in which keyframe extraction method is utilized before extracting spatial features. The temporal features are extracted using attentive correlated temporal feature (ACTF) which uses long short-term memory (LSTM) for deep features. The spatial and temporal features are fused and classified using multi-support vector machine (multiSVM) classifier. Experiments are done on HMDB51 and UCF101 datasets. The results of the proposed method are compared with recent methods in terms of accuracy. The proposed method is proven to work better than other methods by achieving an accuracy of 96% for HMDB51 dataset and 98% for UCF101 dataset.

KEYWORDS

Bidirectional LSTM, Convolutional Neural Network (CNN), Correlation, Keyframe, Long Short-Term Memory (LSTM)

1.INTRODUCTION

In the last half decade, there has been rapid development in the field of Machine Learning. There are lots of components in machine learning that are used to solve many problems. One among them is human action recognition. Deep learning solves several problems in real time applications. It also has its footsteps in recognizing human actions.

The performance of any recognition system depends on whether it is able to extract relevant features. However, extracting useful features is difficult due to a number of complexities. Hence, it is crucial to design any recognition system that can deal with these challenges while preserve categorical information of action classes. Recently, Convolutional Networks (ConvNets) (LeCun et al., 1998) have witnessed great success in classifying images and videos (Krizhevsky et al., 2012). It is also used in human action recognition (Karpathy et al., 2014; Simonyan & Zisserman, 2014). Deep ConvNets have a lot of modeling potential and can develop discriminative representations from raw visual input using large-scale supervised datasets. End-to-end deep ConvNets, unlike image classification, have yet to show a meaningful benefit over traditional hand-crafted features for action recognition.

The massive number of cameras produces huge amount of data that needs to be processed for various reasons. In particular, some videos depict events about people such as their activities and behaviors. To effectively interpret this data requires computer vision algorithms that have the ability

DOI: 10.4018/IJCVIP.296584

to understand and recognize human actions. Its main goal is to locate interesting human action frames in videos that can accurately reflect the various motions in their respective recordings. Several papers based on key frames and video summarizing algorithms (Ioffe & Normalization, n.d.; Karpathy et al., 2014; Krizhevsky et al., 2012; Ni et al., 2015; Simonyan & Zisserman, 2014; Varol et al., 2017) have been provided based on this understanding. Keyframe extraction is also an important task in human action recognition.

This work uses two stream network in greater depth to improve the performance resulting in a novel architecture for human action recognition. In the two stream network, spatial features are extracted using MFNet and temporal features are extracted using LSTM and ACTF. There are three main contributions provided in this work. First, pre-processing methods are used to reduce the burden of deep learning networks. Second, only keyframes are used for spatial feature extraction and ACTF is used for temporal feature extraction. Both significantly increase the performance. Third, the spatial and temporal features are fused to obtain spatio-temporal feature for the whole video.

The remaining of the paper is organized as follows: Section 2 gives the work relevant to our work. Section 3 elaborately discusses the proposed method. Section 4 demonstrates the proposed method with experimental study and ablation study. Section 5 concludes the work and gives future scope.

2.RELATED WORKS

This section discusses related works that are used for comparing the performance of the proposed method. The discussion starts with two stream network, followed by 3D networks and finally discusses recurrent networks. Simonyan et al. (2014) developed a method based on a two-stream network which outperforms previous network architecture of human action recognition. Although this method used temporal information in the video, only short-term movement changes are used, without capturing long-range temporal information of the video.

Inspired by the two-stream hypothesis Simonyan & Zisserman, (2014), two identical stream structures are designed to extract the spatial and temporal features of a video Chen et al., (2019). The extracted features from the two streams are fused using the late fusion strategy and each stream is implemented by an independent ConvNet. The spatial stream works on a single RGB image, and the temporal stream uses a group of consecutive optical flow fields as inputs.

Feichtenhofer et al. Christoph & Pinz, (2016) combined two stream networks with ResNet to form extended ResNet. They have also introduced the residual connection for the appearance stream to increase the interaction between two streams. Temporal Segment Network (TSN) is developed in Wang et al., (2016) to extract the long-range temporal information from the video data. The former papers Christoph & Pinz, (2016); Simonyan & Zisserman, (2014); Wang et al., (2016) use optical flow for temporal feature extraction. To further boost the performance of optical flow estimation, TVNet Fan et al., (2018) is introduced.

There are several researches based on 3D convolution networks. Diba et al. (2017) introduced a Temporal 3D ConvNet (T3D) for action recognition which uses a 3D DesnseNet-based architecture and a new temporal layer Temporal Transition Layer (TTL). This method has a drawback that it uses only spatial features and ignored the temporal features which is more important in action recognition.

A Pseudo-3D Residual Net (P3D ResNet) is developed in Qiu et al., (2017) by reframing 3D convolution layers. They replaced 3D convolutions as 1 x 3 x 3 for spatial features and 3 x 1 x 1 convolutions filters for temporal features. This method has successfully improved the performance of action recognition.

MFNet architecture is introduced in Chen et al., (2018) that divides a complex neural network into an ensemble of lightweight networks also named as fibers. To facilitate information flow between fibers, multiplexer modules are implemented that reduces the computational cost of 3D networks. A temporal feature extraction method named Attentive Correlated Temporal Feature (ACTF) Xu et

al., (2021) is developed by exploring inter-frame correlation within a certain region. This method is better than optical flow-based methods.

Following are some of the papers in human action recognition based on recurrent neural network. Inspired by the success of LSTM network Graves et al., (2013); Vinyals et al., (2015), more researchers used LSTM in action recognition. Donahue et al. (2015) introduced a Long-term Recurrent Convolutional Neural Network (LRCN) which uses the CNN to extract the spatial features and LSTM to extract temporal features. In Veeriah et al., (2015), Differential LSTM is used which added a new gating into LSTM to keep track of the memory states to discover patterns.

Apart from the above categories, a new video representation for action classification is introduced that aggregates local convolutional features across the entire spatio-temporal extent of the video Girdhar et al., (2017). This method integrates two-stream networks with learnable spatio-temporal feature aggregation.

Following are some of the recent works in human action recognition.

In Indhumathi et al., (2022), Adaptive motion ACTF is used for temporal feature extractor. The temporal average pooling in inter-frame is used for extracting the inter-frame regional correlation feature and mean feature. Another model Nasir et al., (2021) is designed to classify human actions by removing redundant frames from videos. In this method, Segments of Interest (SoIs) are extracted. A Neuro Fuzzy Classifier is used at the end for the classification purpose. A novel approach has been introduced for segmented frames and multi-level features sets are extracted Khan et al., (2021). The features are classified by a multi SVM for action identification.

Another method is developed to improve the contrast of video frames using HSI color transformation Afza et al., (2021). Shape, texture and motion features are extracted and fused by a new parallel approach. A new Weighted Entropy-Variances approach is applied to a combined vector and multi SVM is used for classification. A novel 26-layered CNN architecture is designed for complex action recognition Khan, Zhang, Khan et al, (2020). The features are extracted from global average pooling layer and Fully Connected (FC) layer and fused by high entropy based approach. Another feature selection method named Poisson distribution is developed along with Uni-variate Measures (PDaUM).

In Khan, Javed, Khan et al, (2020), fusion of deep neural network and multiview features are done. The best features are selected by higher probability based threshold function. The final feature set is classified using Naive Bayes classifier for final recognition. In Khan et al., (2019), a parallel design utilizing attention-based motion estimation and segmentation module has been developed to avoid the detection of false moving regions. In addition to these contributions, the novel entropy controlled principal components feature selection technique with weights minimization has been implemented to improve the classification accuracy.

3.PROPOSED METHODOLOGY

The proposed method consists of 5 phases: Sequence Reduction, Center Cropping, Keyframe Extraction Feature Extraction and Recognition. In these, the first two phases are pre-processing phases which are used to reduce the computation burden suffered by deep networks.

General Framework

In this work, temporal and spatial features are extracted and processed separately. Figure 1 shows the overall framework of the proposed method. Given an input video as a sequence of frames, the sequences of longer length are removed. To cope with the pre-trained network, the input video sequence should be resized. Instead of meaningless resizing, the frames are cropped at the center. Deep features are extracted using LSTM Vinyals et al., (2015) and MFNet Chen et al., (2018) to produce temporal and spatial features respectively. These features are concatenated and classified using multi SVM classifier which is discussed later in this section.

Sequence Reduction

The input video sequences consist of video of varying sizes. In general, the dataset contains very less sequences of longer length. This can be verified using histogram of the sequence length. Sequences which are much longer than typical sequences in the networks may introduce lots of padding into the training process. Having too much padding can negatively impact the classification accuracy. The longest sequences are removed to improve the accuracy. The histogram of sequence length is shown in Figure 2.

Figure 1. System architecture

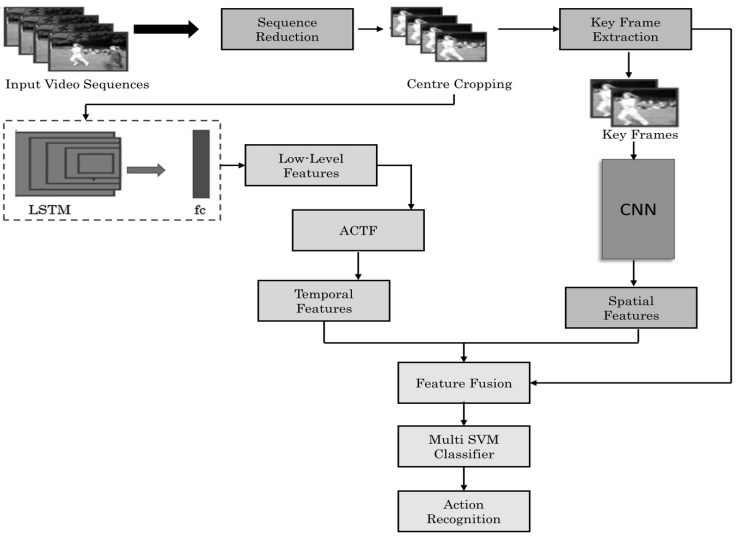
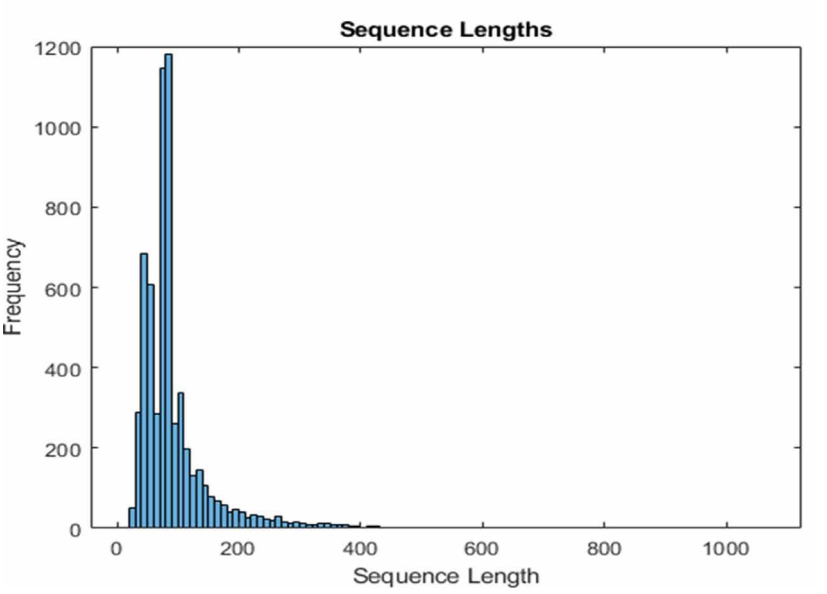


Figure 2. Histogram of sequence length of training sequences



In Fig. 2, it is observed that only few sequences have more than 400 time steps. Hence, the sequences that have more than 400 time steps been removed for training.

Center Cropping

In this work, pretrained deep network models such as LSTM and MFNet are used for feature extraction. The size of these network models is 224 x 224. The input video sequence is reduced to the size of these pre-trained network models. Instead of simply trimming the video size, the unwanted background is cropping so that the center portion matches the size of the pre-trained network models. Thus, the size of the frame in each video sequence after center cropping is 224 x 224. The cropped video sequence is given for temporal feature extraction.

Keyframe Extraction

For spatial feature extraction, only the keyframes are used instead of the whole video sequence. Scene in the video changes from frame to frame Sowmyayani et al., (2014). But all the changes are not visually identified. In this paper, consecutive frames are compared. For evaluating the similarity of frame, Pearson Correlation Coefficient (PCC) Krulikovska & Polec, (2012) is chosen. The value of Pearson Correlation Coefficient can fall between 0(no correlation) and 1(perfect correlation). Correlations above 0.80 are considered as really high and lowest values will be determined as cuts. This threshold is set by several experiments which is discussed in Section IV. The PCC is expressed as

$$PCC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(f\left(i,j\right)-f^{m}\right)\left(f_{p}\left(i,j\right)-f_{p}^{m}\right)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(f\left(i,j\right)-f^{m}\right)^{2}\left(f_{p}\left(i,j\right)-f_{p}^{m}\right)^{2}}} \tag{1}$$

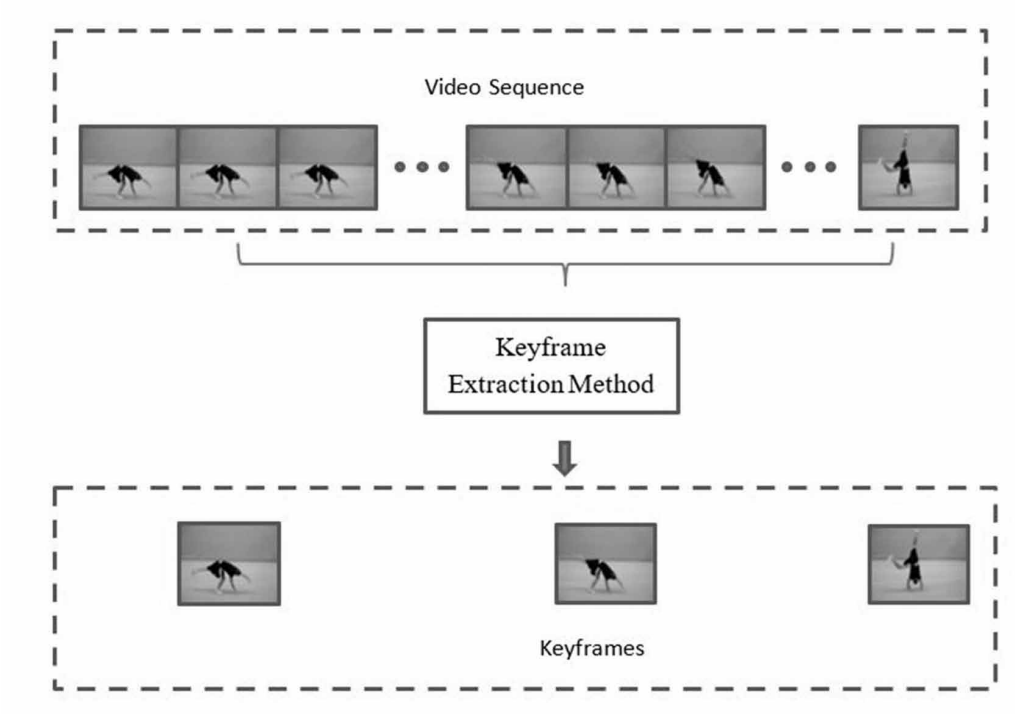
Where f and f_p are two consecutive frames and f^m and f_p^m are the mean intensity of f and f_p respectively. Algorithm 1 gives the keyframe extraction method with its pre-processing phases.

Algorithm 1: Keyframe Extraction

Input: Video sequences
Output: Keyframes
Step 1: maxLength=T1
Step 2: keyframe_threshold=T2
//Sequence Reduction
Step 3: for each video sequence i
Step 3.1 if length(i)>maxLength
set i to null
Step 3.2 end
Step 4: end
//Center Cropping
Step 5: Crop each video at the center
//Keyframe Extraction
Step 6: For each sequence i
Step 6.1: Calculate PCC between i and (i+1) using Eq. (1)
Step 6.2: If PCC>keyframe_threshold
Set i as keyframe
Step 6.3: Else
Go to next frame
Step 6.4: End
Step 7: End

In this algorithm, T1 and T2 are the thresholds for finding the longer sequences and the correlated frame. T1 is found from the histogram of length of the sequences. T2 is calculated by various experiments and found that when T2=0.8, the obtained keyframes are in different scenes. If T2<0.8, some keyframes are in the same scene. If T2>0.8, no keyframe is selected in some scenes. The keyframe extraction method is illustrated in Fig. 2, where few frames from cartwheel sequence consisting 80 frames are shown. The first and last frames are always keyframes in the video sequence. The intermediate frames between two keyframes are called as Group of Frames (GoF). The size of GoF is given to feature fusion phase for further processing.

Figure 3. Illustration of keyframe extraction of cartwheel sequence from HMDB51 dataset



In figure 3, the first keyframe is the representative of first three frames. The fourth frame (not shown in Fig. 2) in the video sequence is the next keyframe. The intermediate frames form a GoF. Hence the size of GoF is 3.

Feature Extraction

The important phase of any recognition algorithm is feature extraction. As the input is video sequence, spatial and temporal features need to be considered. Two different convolutional networks are used to extract spatial and temporal features. This extraction process is discussed in this subsection.

Spatial Feature Extraction:

The extracted keyframes contain only the key evidence of the actions being considered. MFNet is utilized to extract spatial features. As the size of the input keyframes is already resized to MFNet input size, no further augmentation is necessary for using it. Spatial features are extracted from the

last fully connected layer of MFNet. Let $f_{spatial}$ be the feature set obtained from MFNet. The size of this feature set is $t \times C \times H \times W$ where t is of various sizes depends on the number of keyframes in each sequence, C is the number of channels, H and W are the height and width of the feature.

Temporal Feature Extraction:

The temporal feature is extracted from the whole video using LSTM. LSTM uses Googlenet as its backbone network which has the input size as 224 x 224. Hence, data augmentation is not done in this phase. By adding bidirectional LSTM layer to googlenet, LSTM model is created. The features extracted from the last fully connected layer of LSTM are named as low-level features. From these features, ACTF Xu et al., (2021) is calculated using linear and bilinear features for temporal feature extraction. The bilinear operation is used to find interframe correlation within a certain region of successive frames. The ACTF obtained from this phase is $f_{temporal}$. The size of this feature set is $n \times C \times H \times W$ where n is the number of keyframes in each sequence. The temporal feature set of each video sequence is greater than the corresponding spatial feature set.

Feature Fusion and Classification

The dimension of the spatial and temporal feature is of different sizes. Hence concatenating the feature sets has some challenges. In the temporal features, the average of the feature set from one key frame to next keyframe is calculated to match the size of spatial features. Let $f^i_{temporal}$ and $f^i_{spatial}$ be the temporal and spatial feature set of video sequence *i*, where $size\left(f^i_{temporal}\right) > size(f^i_{spatial})$. If the size of the GoF for this video sequence is represented as $V = \left[m_1, m_2, \ldots, m_n\right]$ where *n* is the number of GoFs and obviously it is the size of $f^i_{spatial}$, then Algorithm 3 gives the feature fusion method of single video sequence. The two features are combined to form a Spatio-temporal feature $f^i_{spatio-temporal}$. The set of all Spatio-temporal features obtained in $f\sigma_{temporal}$. Sigma(σ) defines the countable union of extracted spatial and temporal features.

Algorithm 2: Feature Fusion Method

Input: $f^i_{temporal}$, $f^i_{spatial}$, V

Output: $f^i_{spatio-temporal}$

Steps:

Step 1: j=1

Step 2: for every feature *f* in $f^i_{spatial}$

Step 3: Obtain *V[j]* feature set from $f^i_{temporal}$

Step 4: Calculate mean of the obtained set in Step 3, $\sigma_{temporal}$.

Step 5: $f_{spatio-temporal} = \left[f\ \sigma_{temporal}\right]$

Step 6: Obtain $f^i_{spatio-temporal}$ by concatenating $f_{spatio-temporal}$ row-wise.

Step 7: Concatenate

The number of keyframes obtained from keyframe extraction method is given to feature fusion to form Spatio-temporal feature. This feature can be classified using multiSVM Classifier. Algorithm 3 gives the steps for feature extraction, fusion and Recognition.

Algorithm 3: Feature Extraction, Fusion and Recognition

Input: Video Sequences, Keyframes
Output: Accuracy
Steps:
//Spatial Feature Extraction
Step 1: Extract spatial feature FSpatial using MFNet
//Extract temporal feature using LSTM
Step 2: Convert frames to feature vectors
Step 3: Partition the data into training and validation sets
Step 4: Load pretrained Googlenet
Step 5: Add LSTM layers
Step 6: Specify training options
Step 7: Train the network for training data
Step 8: Extract low-level features from the fc of LSTM network
Step 9: Calculate FTemporal by calculating ACTF from the low-level features
//Feature Fusion
Step 10: Call Algorithm 2.
Step 11: Classify using multiSVM Classifier
//Testing
Step 12: Repeat Steps for validation set
Step 13: Calculate accuracy of validation set.

4.EXPERIMENTAL ANALYSIS

This section is discussing the following topics: Dataset, Performance metrics, hyper parameters of deep learning models, comparison of recent methods, ablation study on ACTF and feature fusion.

Dataset Description and Performance Metrics

This performance of the proposed method is evaluated on some datasets: UCF101 Soomro et al., (2012) and HMDB51 Jhuang et al., (2011). The details of the dataset are discussed below. The efficiency of the proposed method is proved by comparing the results of the state-of-the-art methods.

The UCF101 dataset contains 13320 action videos and there are 101 categories. HMDB51 dataset contains 6766 videos and there are 51 categories. UCF101 videos has a fixed resolution of 320×240 which shows a large diversity with respect to the action classes, background variations, illumination, motion of the camera and changes in the viewpoint, as well as the appearance of the object, scale and pose. The challenging dataset with higher intra-class variations is HMDB51 which leads to many errors during both training phase and testing phase. The properties of the both UCF101 and HMDB51 datasets are shown in Table 1. The dataset is splitted into 80% training and 20% testing.

Table 1. Properties of datasets

Property	UCF101	HMDB51
Number of Action Classes	101	51
Number of Video Clips	13320	6766
Resolution	320×240	320 x 240
Frame Rate	25 fps	30 fps
Total Duration	1600 mins	-
Codec	DivX	DivX
Format	AVI	AVI

The accuracy of the proposed system means the accurate segmentation of the pixels and it is calculated as follows:

$$Ac_r = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100 \tag{9}$$

Where Ac_r is the accuracy rate, T_p is the True Positive rate, T_n is the True Negative rate, F_p is the False Positive rate and F_n is the False Negative rates.

Hyperparameters of Deep Learning Models

The proposed method is executed in Nvidia Titan X GPU. Table 2 shows the hyper-parameters used in MFNet and LSTM. For both the models, mini-batch size is set to 16. The stochastic adam optimizer is used in both models.

Table 2. *Hyper-parameters of deep learning models used in proposed method*

Model	MFNet	LSTM
Mini Batch Size	16	16
Dropout	0.8	0.5
Learning Rate	0.001	0.0001
Gradient Threshold	-	2

Comparison of Proposed Method with Recent Methods

The accuracy obtained by the proposed method and some recent methods for both datasets is shown in Table 3. The methods used for comparison is discussed in Section 2.

Table 3. *Comparison of proposed method with recent methods*

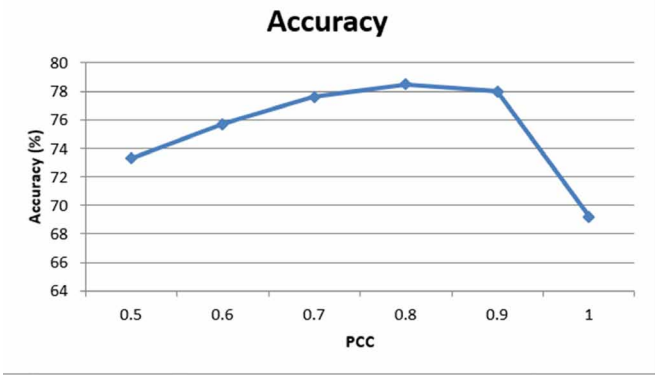
Method/Database	Accuracy (%)	
	UCF101	HMDB51
ST- Isomorphism Chen et al., (2019)	93.9	65.3
ST Heterogeneity Chen et al., (2019)	94.4	67.2
ST- Resnet Christoph & Pinz, (2016)	93.4	66.4
TSN Wang et al., (2016)	94	68.5
TVNet Fan et al., (2018)	95.4	72.5
TSN+ T3D Diba et al., (2017)	93.2	63.5
P3D+IDT Qiu et al., (2017)	93.7	-
MFNet Chen et al., (2018)	96.0	74.6
MFNet-ACTF Xu et al., (2021)	96.3	76.3
ActionVLAD Girdhar et al., (2017)	92.7	66.9
Proposed Method	96.9	78.5

From the Table 3, it is observed that all the methods achieve above 92% for UCF101 dataset and above 63% for HMDB51dataset. The proposed method got 96.9% and 78.5% accuracy for UCF101 and HMDB51 dataset respectively. It is also observed that MFNet-ACTF method achieved a higher accuracy of 96.3% and 76.3% for UCF101 and HMDB51 datasets respectively when compared to other methods. But our proposed method achieves still more accuracy than MFNet-ACTF method. This is due to the fact that we are including the spatial features by extracting keyframes. The proposed method also works better for large-scale dataset than small-scale dataset.

Ablation Study

In this section, we justify our proposed architecture of feature fusion through ablation study. The threshold of PCC in keyframe extraction is set by some experiments. For this study, only HMDB51 dataset is used. The PCC is set from 0.5 to 1. The chart showing the accuracy for various threshold levels are shown in Figure 4.

Figure 4. Line Chart Showing the Accuracy for Various PCC Values



From Fig. 4, the accuracy of the proposed method reaches its maximum when PCC value is set to 0.8. After setting the threshold, we first examine the proposed design without including temporal feature extraction. Experiment is done with the features extracted from MFNet alone (i.e., spatial features). We then examine the inclusion of temporal features by adding LSTM for temporal feature extraction. Finally, we added the ACTF feature in temporal feature set to know its efficiency. The results of all the three designs are shown in Table 4.

Table 4. *Comparison of various design that use partial feature set*

Method	Accuracy (%)	
	UCF101	HMDB51
MFNet	93.8	73.8
LSTM+MFNet	95.6	75.3
LSTM+ACTF	94.66	74.67

From Table 3 and 4, it is studied that the other designs achieve accuracy lesser that the proposed method. The next focus goes to the fusion of temporal feature to the spatial feature. For this, we have used mean and maximum pooling. The results of these two fusions are shown in Table 5.

Table 5. Comparison of fusion of temporal features

	Accuracy (%)	
	UCF101	HMDB51
Mean Pool	96.9	78.5
Max Pool	96.4	78.1

From Table 5, it is studied that the mean pool achieves higher accuracy than max pool. Thus, the design of the proposed method includes LSTM+ACTF for temporal feature extraction, MFNet for spatial feature extraction and mean pool for feature fusion.

5.CONCLUSION

This work extracts spatio-temporal features for human action recognition. Temporal low-level features are extracted from LSTM model from which ACTF is extracted based on correlation. Similarly, spatial features are extracted from keyframes using CNN. The mean pool of temporal features is concatenated with spatial features to form spatio-temporal features. Finally, the features are classified using multiSVM classifier. The proposed method proves its efficacy by experimenting on HMDB51 and UCF101 datasets. The proposed method has achieved an accuracy of 96.9% and 78.5% for UCF101 and HMDB51 dataset respectively. There is a little increase in accuracy obtained by the proposed method when compared to other methods. The proposed method has achieved higher accuracy for large-scale dataset than small-scale dataset. In future, this method can be enhanced to work for small-scale datasets as well.

REFERENCES

Afza, F., Khan, M. A., Sharif, M., Kadry, S., Manogaran, G., Saba, T., Ashraf, I., & Damaševičius, R. (2021). A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. *Image and Vision Computing*, *106*, 104090. doi:10.1016/j.imavis.2020.104090

Chen, E., Bai, X., Gao, L., Tinega, H. C., & Ding, Y. (2019). A spatiotemporal heterogeneous two-stream network for action recognition. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 57267–57275. doi:10.1109/ACCESS.2019.2910604

Chen, Y., Kalantidis, Y., Li, J., Yan, S., & Feng, J. (2018). Multi-fiber networks for video recognition. In *Proceedings of the european conference on computer vision (ECCV)* (pp. 352-367). Academic Press.

Christoph, R., & Pinz, F. A. (2016). Spatiotemporal residual networks for video action recognition. *Advances in Neural Information Processing Systems*, 3468–3476.

Diba, A., Fayyaz, M., Sharma, V., Karami, A. H., Arzani, M. M., Yousefzadeh, R., & Van Gool, L. (2017). *Temporal 3d convnets: New architecture and transfer learning for video classification.* arXiv preprint arXiv:1711.08200.

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). doi:10.1109/CVPR.2015.7298878

Fan, L., Huang, W., Gan, C., Ermon, S., Gong, B., & Huang, J. (2018). End-to-end learning of motion representation for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6016-6025). doi:10.1109/CVPR.2018.00630

Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., & Russell, B. (2017). Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 971-980). doi:10.1109/CVPR.2017.337

Graves, A., Jaitly, N., & Mohamed, A. R. (2013, December). *Hybrid speech recognition with deep bidirectional LSTM. In 2013 IEEE workshop on automatic speech recognition and understanding*. IEEE.

Indhumathi, C., Murugan, V., & Muthulakshmi, V. (2022). Human Action Recognition Using Spatio-Temporal Multiplier Network and Attentive Correlated Temporal Feature. *International Journal of Image and Graphics, 21*(2).

Ioffe, S., & Normalization, C. S. B. (n.d.). *Accelerating Deep Network Training by Reducing Internal Covariate Shift.* arXiv preprint arXiv:1502.03167.

Jhuang, H., Garrote, H., Poggio, E., Serre, T., & Hmdb, T. (2011, November). A large video database for human motion recognition. *Proc. of IEEE International Conference on Computer Vision, 4*(5), 6.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & 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). doi:10.1109/CVPR.2014.223

Khan, M. A., Akram, T., Sharif, M., Muhammad, N., Javed, M. Y., & Naqvi, S. R. (2019). Improved strategy for human action recognition; experiencing a cascaded design. *IET Image Processing*, *14*(5), 818–829. doi:10.1049/iet-ipr.2018.5769

Khan, M. A., Alhaisoni, M., Armghan, A., Alenezi, F., Tariq, U., Nam, Y., & Akram, T. (2021). Video Analytics Framework for Human Action Recognition. *CMC-Computers Materials & Continua*, *68*(3), 3841–3859. doi:10.32604/cmc.2021.016864

Khan, M. A., Javed, K., Khan, S. A., Saba, T., Habib, U., Khan, J. A., & Abbasi, A. A. (2020). Human action recognition using fusion of multiview and deep features: An application to video surveillance. *Multimedia Tools and Applications*, 1–27. doi:10.1007/s11042-020-08806-9

Khan, M. A., Zhang, Y. D., Khan, S. A., Attique, M., Rehman, A., & Seo, S. (2020). A resource conscious human action recognition framework using 26-layered deep convolutional neural network. *Multimedia Tools and Applications*, 1–23.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Krulikovska, L., & Polec, J. (2012). GOP structure adaptable to the location of shot cuts. *International Journal of Electronics and Telecommunications*, *58*(2), 129–134. doi:10.2478/v10177-012-0018-2

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. doi:10.1109/5.726791

Nasir, I. M., Raza, M., Shah, J. H., Khan, M. A., & Rehman, A. (2021, April). Human Action Recognition using Machine Learning in Uncontrolled Environment. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)* (pp. 182-187). IEEE. doi:10.1109/CAIDA51941.2021.9425202

Ni, B., Moulin, P., Yang, X., & Yan, S. (2015). Motion part regularization: Improving action recognition via trajectory selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3698-3706). doi:10.1109/CVPR.2015.7298993

Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5533-5541). doi:10.1109/ICCV.2017.590

Simonyan, K., & Zisserman, A. (2014). *Two-stream convolutional networks for action recognition in videos.* arXiv preprint arXiv:1406.2199.

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

Sowmyayani, S., Rani, J., & Arockia, P. (2014). Adaptive GOP structure to H. 264/AVC based on Scene change. *ICTACT Journal on Image & Video Processing, 5*(1).

Varol, G., Laptev, I., & Schmid, C. (2017). Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1510–1517. doi:10.1109/TPAMI.2017.2712608 PMID:28600238

Veeriah, V., Zhuang, N., & Qi, G. J. (2015). Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 4041-4049). doi:10.1109/ICCV.2015.460

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164). doi:10.1109/CVPR.2015.7298935

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, October). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Springer. doi:10.1007/978-3-319-46484-8_2

Xu, Y., Yang, J., Mao, K., Yin, J., & See, S. (2021). Exploiting inter-frame regional correlation for efficient action recognition. *Expert Systems with Applications*, *178*, 114829. doi:10.1016/j.eswa.2021.114829

C. Indhumathi is currently working as an Assistant Professor in Manonmaniam Sundaranar University Constituent College of Arts amd Science, Kadayanallur. And also she is pursuing her Ph.D in Manonmaniam Sundaranar University. She completed her M.Phil degree in Madurai Kamaraj University and M.Sc degree in Manonmaniam Sundaranar University. Her research interests include Image Processing and Neural Network.

V. Murugan is working as an Assistant Professor in the Department of Computer Science, Manonmaniam Sundaranar University Constituent Arts & Science College, Kadayanallur. He has completed his Ph. D in Computer Science & Engineering from the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli in the year 2016.

G. Muthulakshmi is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli. She received her B.E degree in Department of Computer Science and Engineering, PSR Engineering College, Viruthunagar and M.E degree in Computer Science and Engineering from Manonmaniam Sundaranar University. She received her Doctorate in Computer Science and Engineering from Manonmaniam Sundaranar University. She has published papers in many National and International level Journals and Conferences. Her Research Interests are in the field of Digital Image Processing, Pattern Recognition, and Neural Network.