International Journal of Engineering

Journal Homepage: www.ije.ir

Multimodal Spatiotemporal Feature Map for Dynamic Gesture Recognition from Real Time Video Sequences

S. Reddy P., C. Santhosh*

Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur AP, India

PAPER INFO

Paper history:
Received 21 March 2023
Received in revised form 15 May 2023
Accepted 16 May 2023

Keywords:
2D Video Data
3D Video Data
Human Action Recognition
Visual Saliency
Deep Learning

ABSTRACT

The utilization of artificial intelligence and computer vision has been extensively explored in the context of human activity and behavior recognition. Numerous researchers have investigated and suggested various techniques for human action recognition (HAR) to accurately identify actions from real-time videos. Among these techniques, convolutional neural networks (CNNs) have emerged as the most effective and widely used for activity recognition. This work primarily focuses on the significance of spatial information in activity/action classification. To identify human actions and behaviors from large video datasets, this paper proposes a two-stream spatial CNN approach. One stream, based on RGB data, is fed with the spatial information from unprocessed RGB frames. The second stream is powered by graph-based visual saliency maps generated by GBVS (Graph-Based Visual Saliency) method. The outputs of the two spatial streams were combined using sum, max, average, and product feature fusion techniques. The proposed method is evaluated on well-known benchmark human action datasets, such as KTH, UCF101, HMDB51, NTU RGB-D, and G3D, to assess its performance Promising recognition rates were observed on all datasets.

doi: *10.5829/ije.2023.36.08b.04*

1. INTRODUCTION

Human action recognition from 2D RGB videos has been an active area of research, and various solutions have been proposed, ranging from feature-based approaches to machine learning-based models [1]. However, most of these solutions have primarily focused on basic human movements like walking, jumping, and crawling, and the datasets used have been limited to molecular images, 2D videos (offline and/or online), and 3D depth videos. While the 3D depth and 3D mocap data are popular. They are not practical for real-time use, which raises concerns about their feasibility. Consequently, our research is focused on using 2D videos from lab-captured and online action datasets, and we are exploring the use of feature extraction from 2D videos for identifying human actions.

Classifiers such as Support Vector Machine (SVM), graph matching (GM) [2], adaptive graph matching (AGM) [3], adaptive kernel matching (AKM) [4], artificial neural networks (ANNs), Hidden Markov models (HMM), and Convolutional neural networks (CNN) have been employed for modelling spatial, temporal and concatenation of both in RGB videos. However, the impact of CNN's and their derivatives on action recognition models utilizing 2D video data has been significant [5]. Abaei Kashan et al. [6] investigated the strengths of two descriptors, namely local binary pattern (LBP) and Histogram of Oriented Gradient (HOG), to extract local characteristics. The considered descriptors are widely used in computer vision applications. Azimi et al. [7] proposd method for fully automated image segmentation, which involves layer segmentation and the use of a fully convolutional network (FCN) for task-specific segmentation. CNNs are not only capable of recognizing efficiently but also have ability of extracting a set of actions from an input video sequence when provided with the ideal filter length on every convolution layer [8]. Trained human beings are capable of performing highly complex action sequences, which poses a challenge for extracting a multitude of

*Corresponding Author Email: csanthosh@kluniversity.in
(C. Santhosh)

Please cite this article as: S. Reddy P., C. Santhosh, Multimodal Spatiotemporal Feature Map for Dynamic Gesture Recognition from Real Time Video Sequences, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 08, (2023), 1440-1448

variations in such videos captured during live performances. To accurately interpret complex poses, it has become essential to extract spatial and temporal data with precision. In this work the human action recognition from video sequences are identified by the spatial information.Conv-Nets are a powerful tool to recognize human actions from 2D video sequences [9]. Numerous neural networks have been proposed for this task, many of which are regarded as state-of-the-art in terms of performance [10]. However, some of these networks suffer from drawbacks such as overfitting, inadequate data representation, and suboptimal selection of network architecture [11]. The other challenges include the selection of video frames per class, which in this case, changes depending on the actor performance and the unconstraint nature of videos captured using different imaging sensors.

We propose a multi stream CNN architecture with two streams to address the above challenges effectively. The two streams extract spatial information for classification. The first spatial stream is fed with RGB video frames and the other with graph based visual saliency maps of the action sequences.

Organization of manuscript is as follows: section 1 introduces human recongntion models with the necessity for 3D and 2D video sequences and section 2 deals about the layers on CNN. Whereas the dual stream CNN architecture was illustrated in the section 3 which follows the network architecture and training. Section 5 deals with the results and discussion followed by conclusion and future scope section.

1. 1. Literature Survey Convolution parameters optimization for CNNs, referred as CPOCNN which assigns adaptive upper-bounds of convolution parameters depends on data dimension in current layer and number of remained layers to reach the output layer is propsed by Chegeni et al. [12]. Several fusion methods have been proposed to combine the features extracted from the spatial and temporal streams. The most common methods are early fusion, late fusion, and hybrid fusion. Early fusion combines the features from the two streams at the input level. Late fusion, on the other hand, combines the scores obtained from the two streams at the output level [13]. Hybrid fusion combines both early and late fusion methods to combine the features from the two streams. Recently, several studies have proposed novel methods to improve the performance of the two-stream CNN approach [14]. Scherer et al. [13] proposed a Spatial Temporal Inception module that combines the spatial and temporal streams at different levels of the network. This method has shown to improve the performance of the two-stream CNN approach on the UCF101 [15] and HMDB51 [9] datasets. Liu et al. [16] proposed a Multi-Scale Temporal Attention module that learns the importance of different temporal scales of the video. This method has shown to improve the performance of the two-stream CNN approach on the UCF101, HMDB51, and Kinetics datasets.

Spatio-Temporal ConvNet is a deep learning architecture that processes both spatial and temporal information to classify the action. This architecture consists of 3D convolutional layers that capture the temporal dynamics of the video [17]. Two-Stream ConvNet is an architecture that processes spatial and temporal information separately. The spatial stream processes the raw frames of the video and extracts spatial features. The temporal stream processes the optical flow, which represents the motion information between consecutive frames, to extract temporal features [18]. The two streams are then fused to classify the action. Two-Stream ConvNet has shown to improve the performance of action recognition compared to using only one stream. Long-Term Temporal Convolutions is a technique that improves the performance of action recognition by capturing long-term temporal information. This technique consists of 1D convolutional layers that operate on the temporal dimension of the video [14].

Transfer Learning is a research problem in machine learning that retainsknowledge obtained from the solution of a problem (source domain) to be applied to different but relatively similar problems (target domain). TL has been the mostpopular approach in CNN models in recent years [19]. The 1D convolutional layers have a larger kernel size than the typical 3x3 kernel used in spatio-temporal convolutions, which allows them to capture longer temporal information [20]. Long-Term Temporal Convolutions have shown to improve the performance of action recognition on datasets that involve long-term action [21]. Temporal Segment Networks (TSN) that aggregates features from multiple segments of the video. TSN uses a multi-scale architecture that processes the video at different temporal scales to capture both short-term and long-term temporal information. TSN has shown to achieve state-of-the-art performance on the Kinetics dataset [22].

2. LAYERS OF CNN

The popularity of CNN has increased due to its ability to process large amounts of data. The convolutional layer is the most crucial component of CNN, where the input is convolved using convolution kernels. These kernels act as filters and are followed by a non-linear activation, as defined in Equation (1). This enables CNN to identify distinct representations in speech or image data:

$$act_{i,j} = f\left(\sum_{k=1}^{K}\sum_{l=1}^{L} w_{k,l} \cdot x_{i+k,j+l} + b\right) \quad (1)$$

where, act(i,j) is the respective activation, the weight matrix of the kernel is denoted with w (k,l) of size k×l.

A small bias value b is added and it passed through a non-linear activation f [23].

Rectified linear units (ReLUs) nonlinear function is shown in Equation (2) and is utilized in the convolutional layers to generate the feature maps.

$$\sigma(x) = \text{maximum}(0,x) \quad (2)$$

In general, more hidden features in the input samples can be extracted as the more convolution kernels are included. In contrast to the regular CNN, the model discussed in this paper substituted a global average pooling (GAP) layer for the fully-connected layer that was previously placed behind the convolutional layer. CNN typically ends with single or multiple fully-connected layers that may transform multi-dimensional feature maps into one-dimensional feature vectors. Since each node present in the fully connected layer is linked to a node in the top layer, the fully connected layer's weight parameters may take up the greatest space. The GAP layer implements a global averaging pooling operation on every feature map, in contrast to the fully-connected layer. The GAP layer has no parameters that can be optimized.

During the training process of a neural network, the weight parameters in the top layer continuously change, leading to a continuous shift in the input data distribution for each layer. This dynamic change in distribution poses a challenge for network training and can impede convergence. To address this issue, a batch normalization layer (BN) is introduced after the Global Average Pooling (GAP) layer. The purpose of the BN layer is to adaptively modify the weight parameters to accommodate the evolving data distribution, thereby aiding in the faster convergence of the model.

The BN layer normalizes and reconstructs the input data on each batch of training samples in order to ensure the stability of the output from the previous layer and to improve the speed and accuracy of training.

A Softmax layer as a classifier plus a fully connected layer make up a output layer of CNN [24]. The fully-connected layer should be added at the end of the model because it has a significant advantage. Each node of the fully connected layer is linked to the nodes of the top layer in order to integrate the features that were extracted from the upper layer. In this way, it compensates for the GAP layer's drawbacks.

The Softmax sits underneath the fully connected layer and it transforms the output of the top layer into a probability vector whose value indicates the max likelihood that the current sample belongs to each class. The output of Softmax is given Equation (3).

$$Soft_i = \frac{e^{y_i}}{\sum_{c=1}^{C} e^{y_c}} \quad (3)$$

where, y is the fully connected layer output, C is the number of classes considered.

3. DUAL – STREAM CNN ARCHITECTURE

An action video sequence comprises both spatial and temporal variations observed across a collection of video frames. The spatial component represents the discrete appearances of objects within the frames, while the temporal component captures the movements exhibited by these objects over time. In this work, the action recognition Convolutional Neural Network (CNN) is designed with two spatial streams, as depicted in Figure 1. These streams are constructed using Conv-Nets with SoftMax layers, and late fusion models are employed to calculate similarity scores. Four late fusion models, namely averaging, maximum, product, and sum, are considered in this study.

The architecture comprises of two streams, each containing 8 convolutional layers, two dense layers, and a SoftMax layer. In order to combine the outputs from the SoftMax layers of both streams, a score fusion model utilizing multiple fusion models has been proposed. The proposed architecture is faster and accurate in identifying complex human actions from videos. The model is investigated on multiple action datasets for checking its persistence to multiple types of input video sequences. Robustness of the proposed model is contemplated against various other CNN architectures to ascertain its usefulness in detecting complex actions.

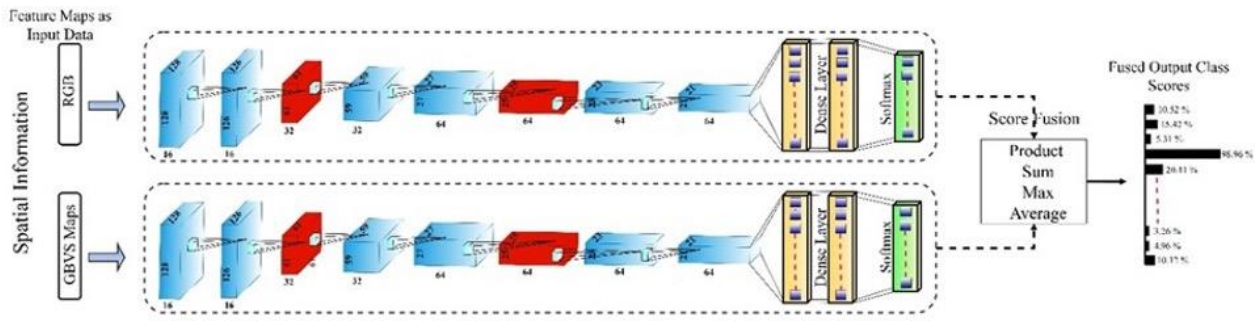

Figure 1. Proposed dual stream architecture

Previous CNN models use RGB sequence as a spatial stream. The main shortcoming is observed that the RGB data in the video sequence poorly represented due to low light indoor environment during a real action performance. To outplay the above shortcoming, we propose to add one more stream which is insensitive to color brightness components in a video sequence.

However, the spatial stream gets graph based visual saliency (GBVS) representing spatial distribution of the action. For the purpose of human action recognition, one of the stream is fed with the GBVS feature maps. The RGB spatial information is totally inconsistent in action sequences captured during a real time performance. The inconsistency appears due to poor lighting and indistinguishable background. To balance this effect, spatial saliency maps are added as an additional stream for the existing RGB stream [23].

To demonstrate the merits of the 2-stream type coding by training a CNN, later the trained CNN is tested to analyze the performance, view invariance and check robustness. We utilized five diverse publicly accessed datasets, KTH , UCF101 [15], HMDB51 [9], G3D [24] and NTU RGB-D [25] for examining the proposed model. The next section briefly enlightens the proposed work and its performance in recognizing human actions from video sequences by performing experiments as an individual streams and 2-streams.

3. 1. Spatial stream-1: RGB stream Conv-Net The proposed architecture incorporates two spatial streams. In the first spatial stream, the RGB frames of an action video sequence are directly utilized. These action sequences are sourced from online datasets, which encompass videos captured under diverse conditions, including complex and simple backgrounds, variations in lighting conditions, and instances of object occlusions.

Figure 2 shows a set of action video frames captured in the wild, which is part of online available datasets.

The frames show a multitude of variations in the action poses with uncontrollable effects from lighting, erratic movements, object occlusion and background variations. The color plays a vital component in identifying a dancer from the background. The first spatial stream is a set of frames in each action class to identify the static pose based on color, texture and shape of the action. Due to large distractions in the data, we propose to double the spatial information model by inducing a second spatial stream based on feature maps extracted from saliency maps.

Figure 2. Real time action performances from online datasets

3. 2. Spatial stream-2 : GBVS stream Conv-Net Graph based visual saliency algorithm is applied to extract the saliency maps from the RGB video frames using graph cut based model. The algorithm was used by Kishore et al. [26] generated the visual saliency maps of video objects. Figure 3 shows the saliency maps created for the sample action video frames from the online datasets.

The saliency maps describe the spatial distribution of the action in the video frame. The second CNN stream takes the video frames shown in Figure 3. This stream of spatial contents is immune to variations such as lighting, color, and background as a reinforcement to the RGB spatial stream. These are a set of low-level features describing the action in an abstract manner.

Incorporating model based spatial representation using GBVS, greatly improves the end-to-end convolution based deep learning methods. When these streams are operated at a time, it also solves the problem of overfitting in the first stream with the weight vectors from the second stream.

4. CNN NETWORK ARCHITECTURE AND TRAINING

The proposed Convolution Neural Network is influenced from the VGG network developed by Simonyan and Zisserman [23] which is extremely deep CNN model that accomplished the state-of-the-art accuracy on Large Scale Visual Recognition Challenge, 2014 classification and localization tasks. VGG net is a densely layered CNN consisting of 16 to 19 weighted layers and a small window of 3×3 aligned along the entire convolution

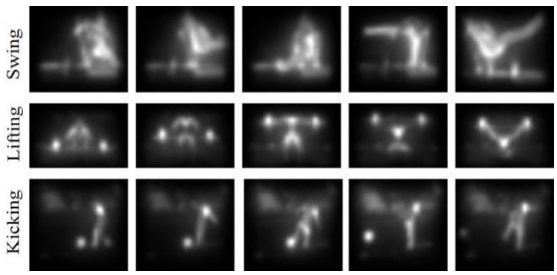

Figure 3. GBVS based saliency maps for action frames of Figure 2

layers. The proposed CNN model is more similar to the indigenous architectures by Ciresan et al. [27] and Dean et al. [28]. Although, the presented Conv-Net architecture is developed from the motivation of VGG net, the depth of weight layers is constrained to 8 and fully connected layers are bounded to 2. Python aided with Keras and TensorFlow libraries are utilized for the construction of this architecture.

The VGG is the origin of every individual stream in the Convnet architecture but limited to the 10 layers. After performing multiple tests using diverse network models such as VGG, Res-Net, Alex-Net and Inception it resulted into the 8 convolution layers preceded by 2 fully connected layers. These are built from the scratch and the overall development of model is done with Keras and TensorFlow in Python 3.6 platform. Additionally, for image classification tasks we trained few pre-trained networks.

4. 1. Dual-stream Conv-Net Training During experiments, the filter configuration of the designated layers varied linearly from 3×3 for initial two layers, 5×5 for the next two, 7×7 and 9×9 for the remaining 4 layers for all sets of training batches. The training algorithm is incepted from Dean et al. [28]. The multinomial logistic regression objective is optimized by training and with the aid of the mini-batch gradient descent when supplied with momentum of 0.9. To normalize the weight decay for 128 frame size at training period, the penalization multiplier is fixed to a range of 2 to 0.0002. Bicubic interpolation is utilized to resize all input frames and they been numbered according to the sequence in each class. For the initial 8 layers the drop out regularization is set to 0.5. When validation accuracy stabilized at a certain constant, the initial learning rate is positioned to 0.02 and declined by a factor of 10. The learning rate plummeted for three times as result it altered from 0.02 to 0.005 to 0.001 and the training has been paused after 10000 epochs i.e., 311.25k iterations and for every 1480 epochs i.e., 150k iterations the learning rate got plunging down. Whereas, when trained with alternate datasets consisting of 500 input size the training terminated at 4800 epochs i.e., 100.125k iterations and learning rate dropped down for every 1220 epochs (10.517k iterations).

The nets required less epochs due to medium scale frame size. In addition, the frame size was raised to the 224×224 in the anticipated labels with negligible or no advancement. In every layer the weights are assigned arbitrarily and the gaussian distribution function and variance are set to zero mean and 0.01 respectively for every layer. The filter outputs from 8 convolutional layers in both the streams are conceptualized in Figure 4.

4. 2. Training Batch Index In order to validate the proposed CNN architecture, we conducted experiments using several 2D action datasets, including

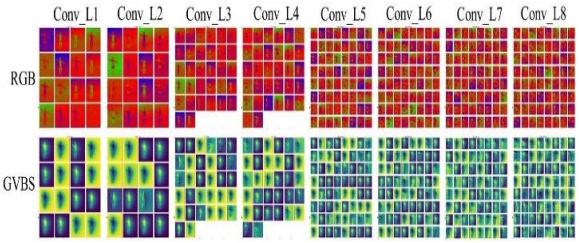

Figure 4. Feature maps visualization of various Convolutional layers

KTH, UCF101, HMDB51, G3D, and NTU RGB-D. Each dataset consisted of 20 different classes with 50 diverse actions, providing a comprehensive evaluation of the developed model. For each dataset, we performed individual training to generate a specific model. Subsequently, the trained model was employed to test the actions within the respective dataset. In order to assess cross-data authentication, we also conducted training with one dataset and testing with another dataset. However, it was ensured that the actions being tested were common across all datasets. HMDB51 and NTU RGB-D datasets were particularly chosen due to their extensive collection of data in various views and diverse subjects. These datasets allowed us to produce comprehensive results and gain a clearer understanding of the proposed method. Throughout this study, the majority of the results were obtained from experiments conducted on these popular datasets, HMDB51 and NTU RGB-D, to provide a detailed analysis of the proposed approach.

4. 3. Testing the Proposed Dual-stream Conv-Net Once the CNN model is trained, it can be used for testing by inputting a batch of 20 action videos. The testing process involves classifying the actions in the videos and assigning corresponding labels. In all the datasets, the 2D video samples have frame sizes of 128x128 pixels. The output of the SoftMax layer is a class score vector, where each element represents the likelihood of the input video belonging to a particular class. Since the proposed Conv-Net consists of two spatial streams, separate class scores are obtained for each stream. To combine the class scores from the two streams, a late fusion approach is applied. This fusion involves using four popular fusion models: maximum, average, sum, and product. Each fusion model generates a single score for each class based on the class scores from the two streams. The network is then tested using multiple experiments focused on human actions. Additionally, pre-trained networks can be utilized for testing by retraining them using available action videos from online datasets. This approach leverages the pre-trained weights and further fine-tunes the network on the specific action videos to improve performance and adapt to the task at hand.

5. EXPERIMENTATION RESULTS AND DISCUSSION

The execution is done with the aid of Keras and TensorFlow toolboxes which can be accessed in python 3.6 substantial adjustments during testing and training. The proposed method is tested on 20 classes of actions from KTH, UCF101, HMDB51, G3D and NTU RGB-D action datasets. Five training models had been produced for the five 5 datasets separately. Performance of each Conv-Net is for a particular dataset is validated with respect to percentage of identification on the complete testing dataset.

5. 1. Evaluation of the Proposed Conv-Net on 2D Action Datasets In this section, the performance of the proposed Dual-stream architecture is evaluated using five publicly available 2D action datasets. Twenty classes with 50 action sequences from the KTH, UCF101, HMDB51, G3D, and NTU RGB-D datasets are selected and labeled appropriately. The CNN architecture remains consistent for each video in the datasets.Each video is segmented into 656 frames, with each frame having a size of 128x128 pixels. Cross-subject testing is conducted on the proposed CNN architecture using 20 test videos from each of the databases. The results of the testing are summarized in the form of a confusion matrix, which shows the classification accuracy and misclassifications for the five test classes. Figure 5 illustrates the confusion matrix, providing a visual representation of the performance of the proposed CNN architecture across the five datasets.

Table 1 gives recognition of the proposed CNN architectures as individual spatial stream and mixed dual-streams across view and subjects. The simulation shows that the proposed architecture utilizes the advantage of both spatial features.

The average fused scores in Table 1, point to the advantage of using multi stream networks compared to single stream. The dual-stream CNN model showed highest recognition rates compared to single stream models. This proves the universal fact in machine learning that wider heterogeneous trainings provide higher recognition.

To visualize the effects of fusion on multi stream CNN architectures, four different score fusion models such as average, maximum, sum and product were tested. Table 2 summarizes the results of the experiment. Product fusion model has proved to produce good score fusions compared to sum, average and max score fusions.

5. 2. Evaluation of the Proposed Conv-Net against other Deep-Nets The dataset is applied on various Conv-Nets such as VGG 16 [23], Alex-Net [29], ImageNet [30], RNN [31], LSTM [32]. At this stage, we study the influence of network model and pre-training on the performance of the task. For this, three cases are formulated: pretrained, pretrained + retrained and training from scratch. The results show training from scratch networks perform better compared to pretrained and pre + retrained. However, training from scratch is computationally intensive due to large data samples fed into the system.

The other model we tested is retraining the pretrained model with new data. Results are summarized in Table 3.

Recognition rates obtained from the state-of-the-art deep nets are compared against the proposed dual-stream architecture on action datasets. The results are averaged over the different testing instances including cross subject, cross view and cross datasets. The recognition rates are tabulated in Table 4.

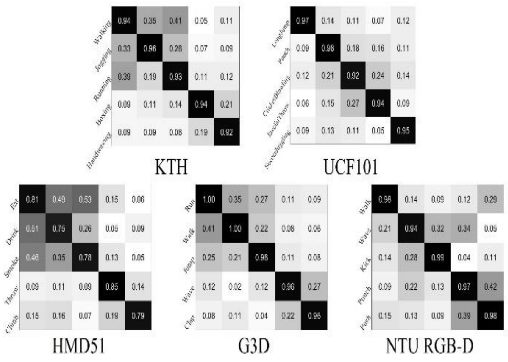

Figure 5. Confusion matrices obtained for few test samples from each dataset on the proposed 2 stream CNN

TABLE 1. Comparison of recognition rates obtained in different individual streams and multi streams of proposed CNN

Dataset	Recognition Rates (%)		
	Spatial Stream		Dual-stream
	RGB	GBVS	RGB + GBVS
KTH	85.37	83.47	92.99
UCF101	89.51	88.25	91.35
HMDB51	66.55	60.18	90.85
G3D	87.43	83.41	92.87
NTU RGB-D	90.29	85.51	94.81

TABLE 2. Testing the score fusion models in 2 Stream networks

Dataset	Score Fusion Method			
	Sum	Max	Average	Product
KTH	92.43	88.57	90.91	92.89
UCF101	95.02	90.12	92.46	91.35
HMDB51	75.63	68.25	70.59	90.45
G3D	89.26	87.58	88.92	92.87
NTU RGB-D	94.59	88.92	91.26	93.81

TABLE 3. Gives the recognition on human action datasets with various Deep Nets

Training Type	Architecture	Action Data				
		HMDB51	**NTU RGB-D**	**KTH**	**UCF101**	**G3D**
Pretrained	VGG 16	83.74	84.45	83.39	83.91	83.73
	AlexNet	77.91	81.75	81.05	79.73	81.08
	ImageNet	79.29	80.78	80.91	80.07	79.56
Pretrained + Retrained	VGG 16	87.13	88.01	87.17	86.93	86.92
	AlexNet	81.52	85.31	84.23	82.56	84.29
	ImageNet	82.85	84.34	82.91	83.49	81.75
Training from Scratch	RNN	85.25	86.29	86.04	85.72	85.82
	LSTM	85.81	87.21	86.29	85.97	86.01
	Ours	90.45	93.81	92.89	91.35	92.87

All architectures are realized with Keras and Tensorflow. The proposed CNN architecture gives highest recognition on both the action datasets, which is due to inclusion of multiple spatial streams for decision making. The above presented results show that the proposed dual-stream CNN gives consistent performance for each of the action with cross view and cross subject variations compared to single stream nets.

5. 3. Evaluation of the Proposed Conv-Net against other Multi Stream Deep-Nets Finally, human action video data is inputted to various multi stream architectures and the average recognition rates were calculated with different late fusion rules.

Table 5 reports the results, showing the proposed multi stream architectures provide better recognition compared other deepNets in literature. The average recognition of our nets touched 94%.

TABLE 4. Comparison of recognition rates for action datasets with different deep learning architectures

Architecture	Action Datasets				
	HMDB51	**NTU RGB-D**	**KTH**	**UCF101**	**G3D**
VGG-VD16 [24]	84.73	89.72	87.46	85.36	87.16
AlexNet [27]	79.99	84.51	83.29	80.99	82.62
ImageNet [28]	80.52	82.96	81.17	81.05	80.94
CNN [9]	85.59	88.95	90.74	86.46	91.25
Proposed	**90.45**	**93.81**	**92.89**	**91.35**	**92.87**

TABLE 5. Performance of multi-stream nets on NTU RGB-D data

Technique	Feature sets	Recognition Rate (%)
Two-stream CNN (fusion-SVM) [9]	Opticalflow + RGB	88.30
Two-stream CNN (fusion-Averaging) [9]	Opticalflow + RGB	86.70
Spatio-temporal ConvNet (Slow Fusion) [17]	High resolution + Low resolution RGBs	89.60
Two-stream ConvNet + LSTM [18]	Opticalflow + Raw RGB frames	88.80
Long-term temporal convolutions (LTC) [21]	MPEG flow + RGB	91.60
DSSCA-SSLM [32]	Depth Maps + RGB	74.56
c-ConvNet [22]	Depth Maps + RGB	89.09
Proposed (Spatial Stream-1)	**RGB**	**85.06**
Proposed (Spatial Stream-2)	**GBVS**	**86.13**
Proposed (Spatial dual-Stream)	**RGB + GBVS**	**94.81**

This is due to the unique architecture having multiple streams in spatial features. Each stream identifies a set of features which are locally crafted by the uniqueness of that algorithm. RGB frame gives static distribution of color brightness in the spatial domain which is supplemented with object action distribution in space to enhance the effect of filters to identify a complex human action correctly.

6. CONCLUSIONS AND FUTURE SCOPE

This work proposed and presented a dual-stream architecture for recognizing complex human actions from 2D videos sequences. The proposed two-stream ConvNets separate spatial information in videos with two streams for each of the spatial data. The two spatial streams are RGB action frames and GVBS based spatial saliency maps. The outputs of 2 SoftMax layers are score fused to generate similarity score. The results showed that the proposed multi streams can recognize human actions accurately and are robust to changing backgrounds in unconstraint videos. Extensive testing proves that the proposed two-stream ConvNets can handle a variety of 2D video data with ease producing consistent outcomes. The average recognition rate on the entire datasets for the proposed two-stream ConvNet is around 94.81%.The dual-stream architecture for human activity recognition has the potential to be further developed and improved to recognize more complex activities. Incorporating additional modalities, attention mechanisms, multi-task learning, and transfer learning can all help to improve the performance of the model for recognizing complex human activities.

7. REFERENCES

1. Afsar, P., Cortez, P. and Santos, H., "Automatic visual detection of human behavior: A review from 2000 to 2014", ***Expert Systems with Applications***, Vol. 42, No. 20, (2015), 6935-6956. https://doi.org/10.1016/j.eswa.2015.05.023
2. Zhou, F. and De la Torre, F., "Factorized graph matching", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 38, No. 9, (2015), 1774-1789. doi: 10.1109/TPAMI.2015.2501802.
3. Yang, X. and Liu, Z.-Y., "Adaptive graph matching", ***IEEE Transactions on Cybernetics***, Vol. 48, No. 5, (2017), 1432-1445. doi: 10.1109/TPAMI.2015.2501802.
4. Popovici, V. and Thiran, J., "Adaptive kernel matching pursuit for pattern classification", in Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, Acta Press. (2004), 235-239.
5. Liu, H., Ju, Z., Ji, X., Chan, C.S., Khoury, M., Liu, H., Ju, Z., Ji, X., Chan, C.S. and Khoury, M., "A view-invariant action recognition based on multi-view space hidden markov models", ***Human Motion Sensing and Recognition: A Fuzzy Qualitative Approach***, (2017), 251-267. doi: 10.1109/TPAMI.2015.2501802.
6. Abaei Kashan, A., Maghsoudi, A., Shoeibi, N., Heidarzadeh, M. and Mirnia, K., "An automatic optic disk segmentation approach from retina of neonates via attention based deep network", ***International Journal of Engineering, Transactions A: Basics***, Vol. 35, No. 4, (2022), 715-724. doi: 10.5829/IJE.2022.35.04A.11.
7. Azimi, B., Rashno, A. and Fadaei, S., "Fully convolutional networks for fluid segmentation in retina images", in 2020 International Conference on Machine Vision and Image Processing (MVIP), IEEE. (2020), 1-7.
8. Srihari, D., Kishore, P., Kumar, E.K., Kumar, D.A., Kumar, M.T.K., Prasad, M. and Prasad, C.R., "A four-stream convnet based on spatial and depth flow for human action classification using rgb-d data", ***Multimedia Tools and Applications***, Vol. 79, (2020), 11723-11746. doi: 10.1109/TPAMI.2015.2501802.
9. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T., "Hmdb: A large video database for human motion recognition", in 2011 International conference on computer vision, IEEE. (2011), 2556-2563.
10. Längkvist, M., Karlsson, L. and Loutfi, A., "A review of unsupervised feature learning and deep learning for time-series modeling", ***Pattern Recognition Letters***, Vol. 42, No., (2014), 11-24. doi: 10.1109/TPAMI.2015.2501802.
11. Simonyan, K. and Zisserman, A., "Two-stream convolutional networks for action recognition in videos", ***Advances in Neural Information Processing Systems***, Vol. 27, (2014).
12. Chegeni, M.K., Rashno, A. and Fadaei, S., "Convolution-layer parameters optimization in convolutional neural networks", ***Knowledge-Based Systems***, Vol. 261, (2023), 110210. https://doi.org/10.1016/j.knosys.2022.110210
13. Scherer, M., Magno, M., Erb, J., Mayer, P., Eggimann, M. and Benini, L., "Tinyradarnn: Combining spatial and temporal convolutional neural networks for embedded gesture recognition with short range radars", ***IEEE Internet of Things Journal***, Vol. 8, No. 13, (2021), 10336-10346. https://doi.org/10.1162/neco.1997.9.8.1735
14. Savadi Hosseini, M. and Ghaderi, F., "A hybrid deep learning architecture using 3d cnns and grus for human action recognition", ***International Journal of Engineering, Transactions B: Applications***, Vol. 33, No. 5, (2020), 959-965. doi: 10.5829/ije.2020.33.05b.29.
15. Soomro, K., Zamir, A.R. and Shah, M., "Ucf101: A dataset of 101 human actions classes from videos in the wild", arXiv preprint arXiv:1212.0402, (2012).
16. Liu, H., Zhou, A., Dong, Z., Sun, Y., Zhang, J., Liu, L., Ma, H., Liu, J. and Yang, N., "M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar", ***IEEE Internet of Things Journal***, Vol. 9, No. 5, (2021), 3397-3415. https://doi.org/10.1162/neco.1997.9.8.1735
17. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L., "Large-scale video classification with convolutional neural networks", in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2014), 1725-1732.
18. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. and Toderici, G., "Beyond short snippets: Deep networks for video classification", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2015), 4694-4702.
19. Zohrevand, A., Imani, Z. and Ezoji, M., "Deep convolutional neural network for finger-knuckle-print recognition", ***International Journal of Engineering, Transactions A: Basics***,

Vol. 34, No. 7, (2021), 1684-1693. doi: 10.5829/IJE.2021.34.07A.12

20. Parvez M, M., Shanmugam, J., Sangeetha, M. and Ghali, V., "Coded thermal wave imaging based defect detection in composites using neural networks", ***International Journal of Engineering, Transactions A: Basics***, Vol. 35, No. 1, (2022), 93-101. doi: 10.5829/ije.2022.35.01A.08.
21. Varol, G., Laptev, I. and Schmid, C., "Long-term temporal convolutions for action recognition", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 40, No. 6, (2017), 1510-1517. https://doi.org/10.1162/neco.1997.9.8.1735
22. Wang, P., Li, W., Wan, J., Ogunbona, P. and Liu, X., "Cooperative training of deep aggregation networks for rgb-d action recognition", in Proceedings of the AAAI conference on artificial intelligence. Vol. 32, (2018).
23. Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, (2014). doi: 10.1109/TPAMI.2015.2501802.
24. Auli, M., Galley, M., Quirk, C. and Zweig, G., "Joint language and translation modeling with recurrent neural networks", in Proc. of EMNLP. (2013).
25. Bloom, V., Makris, D. and Argyriou, V., "G3d: A gaming action dataset and real time action recognition evaluation framework", in 2012 IEEE Computer society conference on computer vision and pattern recognition workshops, IEEE. (2012), 7-12.
26. Kishore, P., Kumar, D.A., Sastry, A.C.S. and Kumar, E.K., "Motionlets matching with adaptive kernels for 3-d indian sign language recognition", ***IEEE Sensors Journal***, Vol. 18, No. 8, (2018), 3327-3337. doi: 10.5591/978-1-57735-516-8/IJCAI11-210.
27. Ciresan, D.C., Meier, U., Masci, J., Gambardella, L.M. and Schmidhuber, J., "Flexible, high performance convolutional neural networks for image classification", in Twenty-second international joint conference on artificial intelligence, Citeseer. (2011).
28. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M.a., Senior, A., Tucker, P. and Yang, K., "Large scale distributed deep networks", ***Advances in Neural Information Processing Systems***, Vol. 25, (2012).
29. Girshick, R., Donahue, J., Darrell, T. and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2014), 580-587.
30. Shahroudy, A., Liu, J., Ng, T.-T. and Wang, G., "Ntu rgb+ d: A large scale dataset for 3d human activity analysis", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2016), 1010-1019.
31. Hochreiter, S. and Schmidhuber, J., "Long short-term memory", ***Neural Computation***, Vol. 9, No. 8, (1997), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
32. Shahroudy, A., Ng, T.-T., Gong, Y. and Wang, G., "Deep multimodal feature analysis for action recognition in rgb+ d videos", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 40, No. 5, (2017), 1045-1058. doi: 10.1109/TPAMI.2017.2691321.

Persian Abstract

چکیده

استفاده از هوش مصنوعی و بینایی کامپیوتری به طور گسترده در زمینه فعالیت های انسانی و تشخیص رفتار مورد بررسی قرار گرفته است. بسیاری از محققان تکنیک های مختلفی را برای تشخیص اقدامات انسانی (HAR)برای شناسایی دقیق اقدامات از ویدیوهای بلادرنگ بررسی و پیشنهاد کرده اند. در میان این تکنیک‌ها، شبکه‌های عصبی کانولوشنال (CNN)به عنوان مؤثرترین و پرکاربردترین شبکه‌های عصبی برای تشخیص فعالیت ظاهر شده‌اند. این کار در درجه اول بر اهمیت اطلاعات مکانی در طبقه بندی فعالیت/عمل متمرکز است. برای شناسایی اعمال و رفتارهای انسانی از مجموعه داده‌های ویدیویی بزرگ، این مقاله یک رویکرد فضایی CNN دو جریانی را پیشنهاد می‌کند. یک جریان، بر اساس داده های RGB، با اطلاعات مکانی از فریم های RGB پردازش نشده تغذیه می شود. جریان دوم توسط نقشه های برجستگی بصری مبتنی بر نمودار ایجاد شده توسط روش (GBVSبرجستگی بصری مبتنی بر نمودار) طراحی شده است. خروجی‌های دو جریان فضایی با استفاده از تکنیک‌های مجموع، حداکثر، میانگین و ترکیب ویژگی محصول ترکیب شدند. روش پیشنهادی بر روی مجموعه داده‌های عملکرد انسانی معیار شناخته شده، مانند KTH، UCF101، HMDB51، NTU RGB-D، و G3D ارزیابی می‌شود تا عملکرد آن ارزیابی شود نرخ‌های تشخیص امیدوارکننده‌ای در همه مجموعه‌های داده مشاهده شد.