New Hybrid Deep Learning Method to Recognize Human Action from Video

Article · August 2021

DOI: 10.26555/jiteki.v7i2.21499

3 authors, including:

Md Shofiqul Islam
Universiti Malaysia Pahang
7 PUBLICATIONS **15** CITATIONS

SEE PROFILE

Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 7, No. 2, August 2021, pp. 306-313
ISSN: 2338-3070, DOI: 10.26555/jiteki.v7i2.21499

New Hybrid Deep Learning Method to Recognize Human Action from Video

Md Shofiqul Islam [1,2], Sunjida Sultana [3], Md Jabbarul Islam [4]
[1] Faculty of Computing, Universiti Malaysia Pahang, 26300, Kuantan, Pahang, Malaysia
[2] IBM Centre of Excellence (Universiti Malaysia Pahang), Cybercentre, Pahang Technology Park, 26300 Kuantan, Pahang, Malaysia
[3] Computer Science and Engineering, Islamic University, Kushtia-7600, Bangladesh
[4] Department of Mathematics, National University, Gazipur-1704, Dhaka, Bangladesh

ARTICLE INFO

Article history:

Received August 08, 2021
Revised August 25, 2021
Accepted September 01, 2021

Keywords:

Convolution;
Deep learning;
Video;
Video action;
Video Classification;
3DCNN

ABSTRACT

There has been a tremendous increase in internet users and enough bandwidth in recent years. Because Internet connectivity is so inexpensive, information sharing (text, audio, and video) has become more popular and faster. This video content must be examined in order to classify it for different purposes for users. Several machine learning approaches for video classification have been developed to save users time and energy. The use of deep neural networks to recognize human behavior has become a popular issue in recent years. Although significant progress has been made in the field of video recognition, there are still numerous challenges in the realm of video to be overcome. Convolutional neural networks (CNNs) are well-known for requiring a fixed-size image input, which limits the network topology and reduces identification accuracy. Despite the fact that this problem has been solved in the world of photos, it has yet to be solved in the area of video. We present a ten stacked three-dimensional (3D) convolutional network based on the spatial pyramid-based pooling to handle the input problem of fixed size video frames in video recognition. The network structure is made up of three sections, as the name suggests: a ten-layer stacked 3DCNN, DenseNet, and SPPNet. A KTH dataset was used to test our algorithms. The experimental findings showed that our model outperformed existing models in the area of video-based behavior identification by 2% margin accuracy.

Md Shofiqul Islam,
Faculty of Computing, Universiti Malaysia Pekan, 26600, Kuantan, Pahang, Malaysia.
Email: shafiqcseiu07@gmail.com

1. INTRODUCTION

The internet is now widely used by individuals all around the world. Social media plays an important role in information distribution. During the same time period, they also express their feelings on social sites for a specific topic so that other users may rapidly learn what is going on. As a result, user opinions are utilized to estimate public reactions to various moves. However, hiring someone to evaluate people's activities through a plethora of content is extremely tough and time-intensive. The researchers use a machine learning approach to classify video activities in order to assess public views. Video action classification is a type of mining that examines video using video processing and feature identification in order to discover people's points of view by collecting and evaluating social and other subjective knowledge resources. Other methodologies are less trustworthy and successful than deep learning methodology. In terms of precision and performance, deep learning-based methodology currently exceeds previous techniques in the image [1], video [2], audio [3], and text [1][4][5][6] classification.

There are many numbers of applications of video sentiment classification. Crime detection from video of real-time game [7], Video Scenario classification [8], Event or occasion video prediction [9], Animated movie video classification [10], Sport player action recognition [11], Movie classification [12].

Some regularly used algorithms are both supervised and uncontrolled, such as SVM and CNN. Throughout most of the action recognition, there is also a range of options (LSTM, GRU, etc.). This introduction part displays the most extensively used video categorization method, as well as its operating mechanism, application benefits, and drawbacks. A method for identifying video using Nave Bayes and a video classification dictionary. If the independence predictors statement is correct, a classifier based on Naive Bayes functions outperforms other models. Another method is to use the internet of video classification to find hateful speech [11]. Another task was to use the SVM for classification with data pilot and weighted production to improve classification performance. When the target classes are overlapping, the SVM algorithm does not perform effectively.

For video labeling, K means are utilized in a variety of ways. Peng [8] recently completed another video categorization task. By segmenting the video, this method is utilized to extract visual features from it and exchange visual feature materials. The typical k-means aggregation approach improves the original grouping values of labialized video samples. When we hold k smalls, K-means execution is usually quicker than hierarchical clusters. The disadvantages are that K-value is difficult to estimate, the K neighbor (KNN) approach is simple and straightforward to use for categorization and extraction, and HMM (Hidden Markov Model) is employed. In the case of R-CNN and HMM methods of real-time video, a new method to the study of kid facial speech [13]. Gains from the HMM method various-length input is a simple generalization for sequences, and quick learning procedures through raw sequence data may occur explicitly. The disadvantages of HMM include the following: HMMs have a lot of unstructured requirements, and they can't rely on hidden states to solve them. Work demonstrates that 3D CNN is most suitable for video classification, as well as analyzing its outcomes under the title of numerous levels pipeline template-based designs to boost the entire 2-D and 3-D CNNs on FPGA [14]. Long-term model RNNs explicitly transfer time variables to video frames of varying lengths. The RNN achieves this by creating networks with iteration that allows information to persist [15]. This loop shape will be used by the neural network to grasp the input series. This is how the RNN works. Wherever we require meaning, RNN can help us out with past comments. There are two forms of LSTM in RNN, as well as one type of GRU. In sporting video sequences with SIFT characteristics, an RNN with neurons in LSTM is being learned [16]. Baccouche's work has long been admired for its steadiness. With the development of deep learning algorithms and models, function extraction becomes automatic. RNN could be improved by using backpropagation. At the conclusion of the day, new slices of footage with improved fidelity were created utilizing 2D Gated Bidirectional Neural Networks for the recognition of violence. Other strategies are less dependable and efficient than deep literacy [4]. This method of learning is more efficient.

2. RELATED STUDY

In fact, leveraging temporal data to the convolutional process is a straightforward approach to action analysis using deep learning. Huang et al. [17] proposed a 3DCNN that uses 3D convolutional layers to get features across spatial and temporary dimensions in order to collect spatiotemporal information across neighboring frames. Even though the 3D CNN is simple, it creates a significant number of parameters, increasing the network's complexity. Due to its thick connections, DenseNets [18] has recently piqued the interest of computer vision scientists. This network has demonstrated outstanding results in natural picture categorization challenges. In a feed-forward way, the system enables any layer to any other layer. During training, this dense connection structure can help with gradient flow. The dense connection of DenseNets decreases the repetitive calls of intermediary parameters, which can reduce a significant set of network parameters and thereby reduce the network's complexity, as compared to the 3D convolutional neural network. DenseNets, on the other hand, has only been used in the context of photos; they haven't fared well in the field of video. This criterion may have an impact on the recognition accuracy of pictures or sub-images of any size or scale. Using the spatial pyramid pooling (SPP) model, the fixed-size constraint can be abolished in this method. Picture cropping is decreased as a result of the application of SPP in the image field, resulting in less loss of image information. Is it possible for the video field to help with such information loss? What effect will implementing SPP to the video world have?. We present a new network model based on video action recognition that takes into account the benefits and drawbacks of both 3D CNN and DenseNet networks, as well as the SPP network model.

To compare the methods, we've included the most relevant papers in video action classification. A method for video action detection has been developed that uses a 3DCNN model with dense pooling and achieves 91.97 % accuracy [19]. This approach is suitable for long videos, but it is also suitable for short video analysis. SVM classification techniques were used to recognize 2391 scenes of six human behaviors acted by 25 participants in four different contexts [20]. With 71.7 % accuracy, this technique performed well in the early stages of video classification. An approach for video recognition using a spatially windowed data algorithm

[21]. This approach achieved an accuracy of 81.2 %, but feature handling for multiscale needs to be improved. An unique 3D CNN model for action recognition that uses high-level characteristics to regularize the outputs and combines the predictions of several independent models [17]. The accuracy of this technique was 90.5 %, however data labelling and preprocessing should be enhanced for even better results.

We suggest a novel technique to address some of the shortcomings of the previous one. A new network design based on 3D convolutional networks, densely CNN, and the SPP network is proposed in this research. "Stacked 3D-DenseNet-SPP" is how we refer to it. As the major network structure in our architecture, we use DenseNet. We can accomplish the goal of merging DenseNet with such a ten stacked 3D CNN by has component information indeed to the DenseNet. As a result, when compared to a 3D convolution network alone, this not only makes full more use temporal and spatial information but also lowers gradient diminishing and minimizes the network's training parameters to some amount. The spatial pyramid pooling architecture is then added to the underlying network to meet any scale/size inputs.

Our main contributions are as follows. The first contribution is to utilize spatial pooling with normalization in 3D-CNN. The next contribution is to select and handle important features. The third contribution is to develop a new hybrid deep learning-based video action classification with ten stacked 3DCNN to get higher accuracy. The last contribution is Performance evaluation and comparison of our proposed method of video action classification to show better performance of our method.

The remainder of this document is organized as follows. The first section provides background information on the video classification technique study. Section 2 contains a critical examination of current relevant research on video action classification, as well as a summary of the problem and our contributions. The primary technique is described in section 3, and the results are described in part 4. Part 5 contains the discussion, while section 6 has the conclusion.

3. METHOD

This section describes the methodology for video action classification. This section includes a description of the different layers to develop the proposed model with main figures, equations, and algorithms. A sequential description of each layer is described below.

The spatial data from the neighbor feature preceding layer can be extracted by a 3D convolution neural network, as well as the temporal features across adjacent frames. In general, a 3D CNN is used to get spatial as well as temporal features from video input in order to recognize actions. Its outstanding performance is due to 3D convolution and 3D pooling operations. Convoluting a 3D kernel to the cube generated by stacking numerous frames integration yields 3D convolution. The local features in the CNN layer are coupled to numerous consecutive frames in the preceding layer using this method, allowing motion information to be captured. We typically need to add a 3D max pooling with 3D convolution layers to minimize the network parameters and size of the video representations. The operation of 3D pooling is as simple as adding the spatial ordering to 2D pooling. 3D pooling procedure in a 3D convolutional neural network can decrease computation and overfitting.

We introduce a new deep learning architecture, a 3D DenseNet with the spatial pyramid pooling (3D-DenseNet), that can work on classification tasks and is influenced by the hierarchical feature pooling model, which is used to photographs of random size and the growing popularity of densely connected network systems in video classification tasks. To create the Stacked 3D DenseNet, we just use DenseNet as a primary network and 3D pooling with 3D CNN.

Dense connections are used to increase information flow across layers, resulting in fewer network parameters and quick running performance. (The levels are linked and can be accessed directly from one another.) The speed will be faster by less transition among layers).

Assume that xl is the production of the l-th layer. The output xl is calculated in typical feed-forward networks by $x_l = f_l(x_{l-1})$, where f is not a linear modification of the l-th layer. In most cases, especially in deep neural networks, this method cannot prevent gradient disappearance or explosion. The l-th layer, on the other hand, must deliver the feature maps xl to all following layers in the dense interconnections structure. Layer l, in turn, will obtain feature maps from all preceding layers as input: $x_l = f_l([x_0, x_1, \ldots, x_{l-1}])$ as integrated of feature maps generated in layers 0, 1,.., l -1]. The function is denoted by the letter f. The DenseNet structure uses better features with gradient transfer learning by employing dense connections. Data in this study comprises not only spatial but also spatial data.

Batch normalization, ReLU activation, and 3D convolution are the three operations that make up the composite function f. This is used as input of the 3D convolution layer. We use many densely connected dense blocks in the proposed model to make greater use of stacked 3D CNN and 3D pooling. A spatial pyramid pooling layer is employed before the fully linked layer to suit the needs of frames of any size.

Conventional SPP networks are commonly employed in 2D network topologies. Through the SPP structure, the top layer's two-dimensional feature map is produced in parallelism to the linked layer. The 3D feature vector is transferred into the SPP framework in our model. To link appropriately, we expand the typical SPP network topology by adding a temporal dimension. At the very same time, the spatial pyramid pool is conducted only in the spatial architecture to guarantee that the SPP layer is not influenced by temporal information interference. At every layer of the SPP, the time data is guaranteed to be the same, and the spatial features are normally measured using the SPP system. Our approach can be seen as a step forward from the densely connected network model in some ways. In comparison to DenseNet, our model is capable of not only processing 3D visual information but also entering video frames of any size. In contrast to the 3D convolutional neural network, our model contains fewer parameter values. Fig. 1 gives illustrations of the main Stacked 3DCNN model

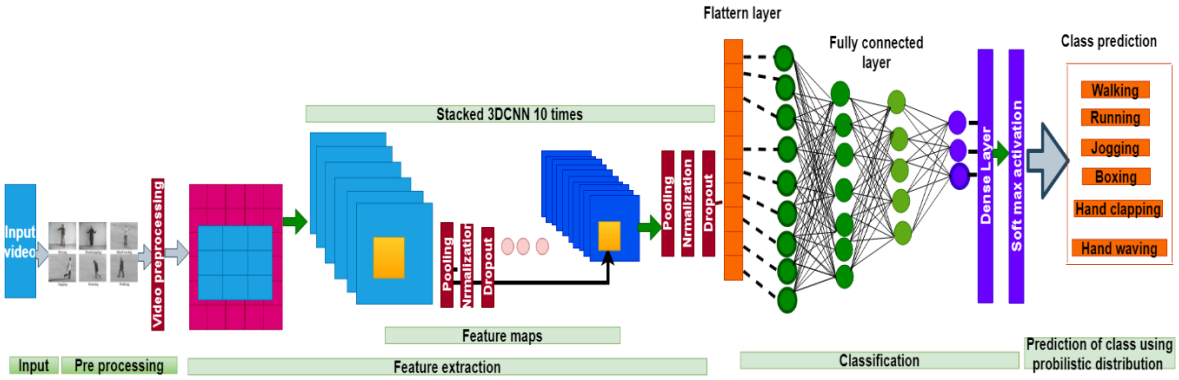

Fig. 1. Main Stacked 3DCNN model

2.1 Algorithm for Model development

This section states the main model for our proposed model. Sequential steps of the model development are given below. This algorithm clearly shows that how our method works and is analyzed. From start to end, each step shows sequential operations done by our proposed method.

Algorithm for model development:
Input: Take input of Filter number
Output: Return full model

Algorithm-Main-Model ()

1. Import related libraries
2. Install libraries
3. Import sequential,3DCNN,Dense,flattern,dropout layer.
4. Set img_rows=160, img_cols=120
5. Set model = Sequential(l)
6. Get img_rows, img_cols
7. define model_N():
8. model.add(Convolution3D(l*2, 9, 7, 3,input_shape=(1, img_rows, img_cols,
9. 43), activation='relu'))
10. model.add(MaxPooling3D(pool_size=(3, 3, 1)))
11. model.add(BatchNormalization(center=True, scale=True))
12. model.add(Dropout(0.5))
13. return model
14. Set n=10
15. Set l=16
16. for i in n:
17. model=model+model_N(l)
18. model.add(Flatten())
19. model.add(Dense(nb_classes,init='normal'))
20. model.add(Activation('softmax'))
21. model.compile(loss='categorical_crossentropy', optimizer='RMSprop', metrics=['accuracy'])
22. Return model

2.2 Data set used in video action classification

In media-related disciplines of research, several authors have spent a lot of work obtaining and labeling video data sets. To assess the performance of our model, we use KTH data. With 2391 video sequences, the KTH dataset contains six categories of human behaviors. All of the segments were shot with a static focus lens at a frame rate of 25 frames per second over unique backgrounds. This video length is 4 seconds and has a resolution of 160 multiplied by120 pixels. More detailed information about KTH data is summarized in Table 1. Source of KTH data link: https://www.csc.kth.se/cvap/actions/. We use 80% data for training and 20% data for validation. The confusion matrix for the KTH actions is given in Fig. 2. This illustrates the correlations among all classes of video action from KTH data.

Table 1. Summary of KTH Video action dataset

Dataset	Year	Total categories	Videos per class	Data type	Total videos
KTH	2004	6	Walking 100, Boxing 100, Handclapping 100, Jogging 100, Handwaving 99, Running 100.	Static	2361

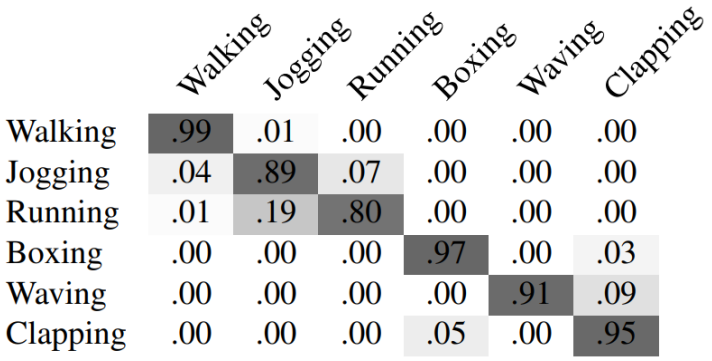

Fig. 2. Confusion matrix for the KTH actions

4. RESULT

This section completely presents the result and analysis. To check the performance of the proposed method, we have compared all the model's accuracy with our model for each class of video from kth data. Here Table 2 shows compared result that shows our model working better than all compared methods. This table shows each class performance on KTH data by different compared methods. From this table, it is clearly shown that our method performs better with the highest average accuracy of 93.8%, and that is about 2% margin accuracy. 3DCNN based approaches are doing better than the SVM algorithm. Among the 3DCNN methods, our stacked-based 3DCNN method is doing better for video action classification from KTH data. Code and data source of this paper: https://github.com/shafiq-islam-cse/Video-action-recognitnion

Table 2. Action recognition performance from KTH data

Method	Features	Walking	Boxing	Clapping	Jogging	Waving	Running	Average
[20]	SVM	83.8 %	97.9%	59.7%	60.4%	73.6%	54.9%	71.7%
[21]	Spatio-Temporal And cuboid	90%	93%	77%	57%	85%	85%	81.2%
[17]	3DCNN	97%	90%	94%	84%	97%	79%	90.5%
[19]	3DCNN Dense pooling	95.4%	89.2%	86.8%	91.6%	92%	97%	91.97%
Our	Stacked 3DCNN	98%	89%	95%	91%	93%	97%	93.8%

Fig. 3 shows the compared result that shows our model working better than all compared methods. Each bar indicates every class performance, and the color is indicated on the side Fig. 3. In this figure, we just showed the performance result on different video actions (walking, Boxing, Handclapping, Jogging, Waving, and Running) recognition from KTH data.

To get deeper details on performance, we run our method for 50 epochs. We track running history to get performance results after every 10 epochs in 50. It is clearly shown in Table 3 that epoch number 50 shows better accuracy than the previous. We have also checked the performance of our model at the different number of epochs. Table 3 shows the performance of our model for the different number of epochs.

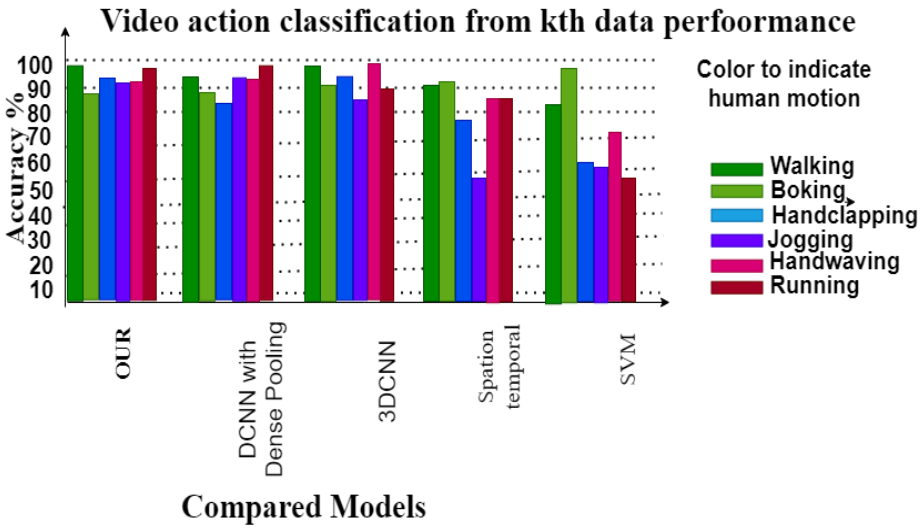

Fig. 3. Compared result

Table 3. Performance on various epochs.

Epoch based performance				
Number of epochs	**Training accuracy**	**Training loss**	**Validation accuracy**	**Validation loss**
10	0.6221	0.8938	0.6583	0.8481
20	0.7411	0.5843	0.6417	0.8484
30	0.7829	0.4761	0.6667	1.2115
40	0.8330	0.3893	0.6333	1.1126
50	0.9381	0.2862	0.6750	1.5378

Fig. 4 and Fig. 5 shows overall training and validation performance, respectively, for our method. The y axis represents number of epochs. In Fig. 4, it is clearly shown that y axis present achieved accuracy at each epoch and in Fig. 5, x axis shows loss.

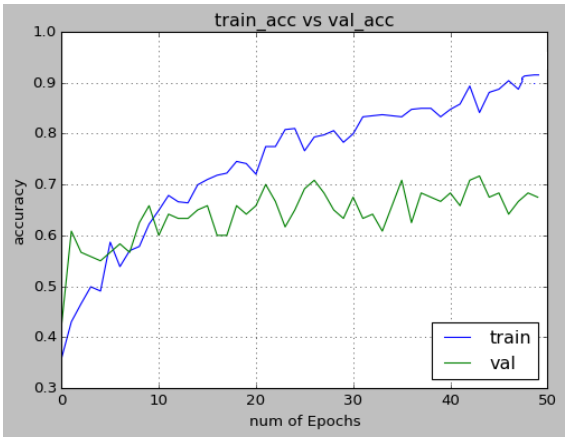

Fig. 4. Model accuracy

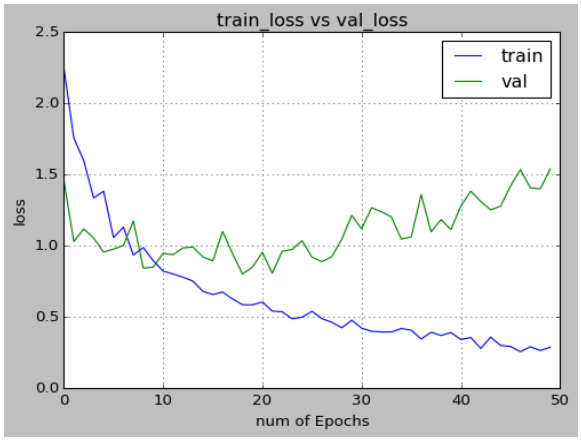

Fig. 5. Model loss

5. DISCUSSION

The results of the experiments revealed that our approach not only enables the training of video frames of any size but also performs better with 2% margin accuracy. For action recognition, we used a mix of the 3DCNN model, Normalization, and DenseNet in this research. There are a variety of additional deep designs, including two-stream ResNet models. This method is frequently employed in the area of pictures and performs well. It would be fascinating to see whether these models could be expanded to include video. This possibility will be investigated further in the next analysis. However, efficient video categorization requires an

understanding of frame attributes. Patterns also help to improve the accuracy of categorization jobs. Another potential difficulty of combining more kinds of videos into the data with more effective and generic capabilities is to investigate camera movement-specific approaches. To classify longer videos, recognize numerous actions in videos, find correlations between videos, and classify multiple object actions in videos. The trend work in action recognition is live steaming gaming video prediction. Both spatial and temporal features can be captured using 3DCNN. However, the fundamental issue with 3DCNN is that its 3D structure is costly to operate.

6. CONCLUSION

The topic of video recognition is covered in this work. We presented a stacked 3D CNN based on SPP to increase the recognition impact and realize video input of any size (Stacked 3DCNN). The model uses a fully connected network (DenseNet) extension, with time information added to all convolutional and pooling layers, as well as a hierarchical spatial structure. The KTH dataset was used to test our model and got 2% margin accuracy than other compared models. Existing approaches include disadvantages such as being unable to handle numerous features at once, taking longer to train with deep learning, being less adaptable with classical machine learning, and having low accuracy when dealing with multilevel video. The researcher's direction and opportunity are to overcome the constraints of video categorization. In the future, we want to be able to classify lengthier videos, recognize numerous actions in videos, find correlations between different videos, and classify multiple object actions in videos. Our future effort in video classification will be live streaming game video prediction.

REFERENCES

[1] I. Khandokar, M. Hasan, F. Ernawan, S. Islam, and M. Kabir, "Handwritten character recognition using convolutional neural network," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1918, no. 4, p. 042152, 2021. https://doi.org/10.1088/1742-6596/1918/4/042152

[2] M. S. Islam, S. Sultana, U. kumar Roy, and J. Al Mahmud, "A review on Video Classification with Methods, Findings, Performance, Challenges, Limitations and Future Work," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 6, no. 2, pp. 47-57, 2020. https://doi.org/10.26555/jiteki.v6i2.18978

[3] L. Fan, Z. Yin, H. Yu, and A. Gilliland, "Using Data-driven Analytics to Enhance Archival Processing of the COVID-19 Hate Speech Twitter Archive (CHSTA)," *preprint*, 2020. https://doi.org/10.31229/osf.io/gkydm

[4] M. S. I. Shofiqul, N. Ab Ghani, and M. M. Ahmed, "A review on recent advances in Deep learning for Sentiment Analysis: Performances, Challenges and Limitations," *COMPUSOFT: An International Journal of Advanced Computer Technology*, vol. 9, no. 7, pp. 3768-3776, 2020. https://ijact.in/index.php/ijact/article/view/1175

[5] M. S. Islam, S. Sultana, U. K. Roy, J. Al Mahmud, and S. Jahidul, "HARC-New Hybrid Method with Hierarchical Attention Based Bidirectional Recurrent Neural Network with Dilated Convolutional Neural Network to Recognize Multilabel Emotions from Text," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 7, no. 1, pp. 142-153, 2021. https://doi.org/10.26555/jiteki.v7i1.20550

[6] M. S. Islam and N. A. Ghani, "A Novel BiGRUBiLSTM Model for Multilevel Sentiment Analysis Using Deep Neural Network with BiGRU-BiLSTM," Singapore, Springer Singapore, vol. 730. pp. 403-414, 2021. https://doi.org/10.1007/978-981-33-4597-3_37

[7] M. Zhen *et al.*, "Learning Discriminative Feature with CRF for Unsupervised Video Object Segmentation," in *European Conference on Computer Vision*, Springer, vol. 12372, pp. 445-462, 2020. https://doi.org/10.1007/978-3-030-58583-9_27

[8] T. Peng, Z. Zhang, K. Shen, and T. Jiang, "Video Classification Based On the Improved K-Means Clustering Algorithm," *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 440, p. 032060, 2020. https://doi.org/10.1088/1755-1315/440/3/032060

[9] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Generation Computer Systems*, vol. 106, pp. 92-104, 2020. https://doi.org/10.1016/j.future.2020.01.005

[10] Z. Li, R. Li, and G. Jin, "Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary," *IEEE Access*, vol. 8, pp. 75073-75084, 2020. https://doi.org/10.1109/ACCESS.2020.2986582

[11] X. Li and S. Geng, "Research on sports retrieval recognition of action based on feature extraction and SVM classification algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 4, pp. 5797-5808, 2020. https://doi.org/10.3233/JIFS-189056

[12] A. Yadav and D. K. Vishwakarma, "A unified framework of deep networks for genre classification using movie trailer," *Applied Soft Computing*, vol. 96, p. 106624, 2020. https://doi.org/10.1016/j.asoc.2020.106624

[13] C. Li, A. Pourtaherian, L. Van Onzenoort, W. T. a Ten, and P. H. De With, "Infant Facial Expression Analysis: Towards A Real-time Video Monitoring System Using R-CNN and HMM," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 1429-1440, 2020. https://doi.org/10.1109/JBHI.2020.3037031

[14] J. Shen, Y. Huang, M. Wen, and C. Zhang, "Towards an efficient deep pipelined template-based architecture for accelerating the entire 2D and 3D CNNs on FPGA," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, pp. 1442-1455, 2019. https://doi.org/10.1109/TCAD.2019.2912894

[15] H. Yang *et al.*, "Asymmetric 3d convolutional neural networks for action recognition," *Pattern recognition*, vol. 85, pp. 1-12, 2019. https://doi.org/10.1016/j.patcog.2018.07.028
[16] A. Kar, N. Rai, K. Sikka, and G. Sharma, "Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3376-3385. https://doi.org/10.1109/CVPR.2017.604
[17] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221-231, 2012. https://doi.org/10.1109/TPAMI.2012.59
[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708, 2017. https://doi.org/10.1109/CVPR.2017.243
[19] W. Yang, Y. Chen, C. Huang, and M. Gao, "Video-based human action recognition using spatial pyramid pooling and 3D densely convolutional networks," *Future Internet*, vol. 10, no. 12, p. 115, 2018. https://doi.org/10.3390/fi10120115
[20] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004.*, vol. 3, IEEE, pp. 32-36, 2004. https://doi.org/10.1109/ICPR.2004.1334462
[21] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, IEEE, pp. 65-72. https://doi.org/10.1109/VSPETS.2005.1570899

BIOGRAPHY OF AUTHORS

Md Shofiqul Islam is doing Masters (Research-based), a student at University Malaysia Pahang (UMP), Pahang, Malaysia. He has completed his B. Sc. in 2014 in CSE from Islamic University, Kushtia, Bangladesh. Now he is a research assistant at University Malaysia Pahang (UMP). He is also a teacher at CSE under the faculty of FST at ADUST university, Dhaka. He is also in the teaching profession since 2015. His research field is Deep learning, Machine learning, Natural Language Processing, Image Processing. He has published a lot of papers in my field. Email: shafiqcseiu07@gmail.com

Sunjida Sultana was completing a master's degree and completed a bachelor's degree from the Department of Computer Science and Engineering, Islamic University, Kushtia-7600, Bangladesh. She is working in the field of image processing, video processing, and text processing. Her email is sunjidasultana51984@gmail.com

Md Jabbarul Islam has completed bachelor's degrees from the Department of Mathematics, National University Gazipur-1704, Dhaka, Bangladesh. He is doing his research work in the field of Graph theory, Statistics, Machine learning, image processing, video processing, and text processing. His email is abduljabbar11061997@gmail.com