2020 International Conference on Computational Science and Computational Intelligence (CSCI) | 978-1-7281-7624-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/CSCI51800.2020.00129

An Activity Recognition Framework for Overlapping Activities using Transfer Learning

Muhammad Bilal
Department of Computer Science,
COMSATS University Islamabad,
Attock Campus, Pakistan

Muazzam Maqsood
Department of Computer Science,
COMSATS University Islamabad,
Attock Campus, Pakistan

Irfan Mehmood
Department of Media Design and Technology, Faculty of Engineering & Informatics, University of Bradford, UK

Mubashir Javaid
Department of Computer Science,
COMSATS University Islamabad,
Attock Campus, Pakistan

Seungmin Rho
Department of Software, Sejong University, Seoul, Korea

Abstract*— Activity recognition is gaining popularity with the increase in digital content. In video data, there is a lot of information hidden that needs to be explored. Human Activity Recognition (HAR) in video streams applies to many areas, such as video surveillance, patient health monitoring, and behavior analytics. *Variation in the environment, view-point changes, occlusion, and illumination are some main challenges in HAR. Among other challenges, there is also a similar activity or overlapping activity issue that has not been explored much in past. Resolving overlapping activities classes issues can be a major contribution to overall Human Activity Recognition. Hand-crafted methods and traditional Machine Learning methods were extensively explored in past. Recently, many Deep Learning-based methods are achieved high accuracy. Convolutional Neural Network (CNN) and 3D CNN methods outperform other methods. In this paper, we proposed a Transfer Learning-based Human Activity Recognition (TLHAR) for video data streams. We used VGG16 and InceptionV3, two pre-trained CNN models, and utilized their prior training knowledge for efficient activity recognition. The purposed system outperformed existing activity recognition methods and showed state-of-the-art accuracy and less computational cost requirements than other techniques by taking the benefits of Transfer Learning.

Keywords—Human Activity Recognition (HAR), Transfer Learning, Convolutional Neural Network (CNN), VGG16, InceptionV3, Video Content Analysis

I. INTRODUCTION

Activity recognition involves the identification of different actions from videos. It can be a single action or a combination of multiple actions. Activity recognition can apply to many domains such as video surveillance, video analytics, human behavior analysis, healthcare, and patient monitoring, human-computer-interaction systems [1]. With the availability of more and more video data, Human Activity Recognition (HAR) has captured extensive attention from research and industrial communities for video analysis [2]. The movement of different body parts helps in identifying human actions in the context of video surveillance. In images, as they are still, it is difficult to recognize and classify the action

An activity can be short or maybe as long as a video. Activity recognition includes capturing Spatio-temporal context across frames. To recognize human activities correctly it is necessary to capture activity in a sequence of frames [3]. There is a local context and also a global context with-respect-to motion in activity recognition. Activity recognition requires activity detection, activity localization, and activity recognition. Activity recognition is the classification of activities into some pre-determined activity classes. Activity detection is to find out whether an interesting activity is present in trimmed/untrimmed videos or not. Activity localization finds the potential proposals in videos when the activity of interest occurs. Detection and analysis of human body parts and also their interaction with other objects and surroundings helps in recognizing actions accurately [4]. An activity tube (i.e., the sequence of action) includes the spatial relationship of activity over time which adds extra complexity for spatial refinement. Handling temporal information becomes difficult for activity classification because it requires temporal contextual information of activities [5].

Spatio-temporal activity recognition is still a challenging task. Variation in new data in video data streams leads to a problem due to its dynamic content that models trained on previous data cannot be applied to new data as it has been changed. Moreover, human action recognition is also challenging due to high dimensional features, occlusion, variation in viewpoint, illumination, motion, and clutter issues [6]. Among other problems expensive high-performance-computation and insufficient annotated data for large videos have become the key challenge. Detection and recognition of human activities in large video streams, differences in performing activities from person-to-person, overlapping classes, and variations in the environment is still a challenging task [7].

Previously different methods were used to solve these problems including optical-flow, spatial-temporal-based activity recognition, 3D depth/sensor data, physics, and skeleton-based methods. Recently Deep CNN-based models showed some progress in action recognition, as they can automatically learn features to distinguish different actions performed in videos[8]. Pre-trained CNN based models have already achieved significant success in performance for video summarization, fire and fog detection, and image retrieval [3, 9].

Some activities have a similar pose in their frames and are difficult to distinguish. For example, rope skipping and jumping for head-shot in football looks similar in their initial phase [10]. In this article, we propose a deep learning-based framework for Human Activity Recognition (HAR) from the data streams of visual surveillance environments. We exploited the idea of Transfer Learning for recognizing Human Activity Recognition (HAR) from videos in the wild. Two pre-trained CNN models: VGG16 and InceptionV3 were utilized for activity recognition. Activities in the UCF-101 data set, with similar actions or overlapping activities

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00129

grouped for correct classification of the actual class. The main contributions to this paper are summarized below:

- Efficiently extract and learn deep features in dynamic video streams containing hidden information.
- Improve activity learning and classification performance on the benchmark dataset.
- Efficiently detect human actions in large video streams and resolve overlapping classes.

The remaining paper is organized as follows: Section II briefly reviews related work about Human Action Recognition (HAR). Section III explains various aspects of the proposed framework. Section IV presents experimentation results and a brief discussion on them. Section V consists of the conclusion and future directions.

II. Related work

Ullah et al. Proposed a human action recognition framework for online video data streams in a non-stationary environment for surveillance. Their proposed system utilizes activations of Fully-Connected Layer (FCL) of a pre-trained VGG16 model for frame-level deep feature extraction. To incorporate temporal changes the model uses Deep Autoencoder (DAE) and Support Vector Machine (SVM) was used to classify action classes. Complex activities that are similar/overlapping actions in them caused a reduction in the recognition score [6]. J. YU et al. presented a Discriminative-Deep-Model (D3D-LSTM) that improves Spatio-temporal action recognition for both single action and recognition of interactions actions. To improve similar actions recognition a feature fusion mechanism was used in a real-time environment. The attention method was used to assign weights to individual frames in real-time. To obtain the best performance parameters an optimization scheme was introduced. To improve performance for long-term video weights were controlled with an attention mechanism by updating the state of the current memory [11]. Liang et al. proposed a multi-model framework for action recognition by considering sub-action extraction and representation learning methods. Long and overlapping activities decomposed into sub-actions with an unfixed number of temporal segmentation of video streams containing sub-actions. Sub-action learning for collaborative local consistency explored the idea of sub-action sharing for high-level discriminative representation Hand-crafted features extraction module could be extended for Deep Neural Network-based of the hybrid framework [4].

Nazir et al. developed a multi-kernel learning mechanism for the temporal Residual. Deep features were represented with the Residual Network by using both motion and appurtenance features. A multi-kernel Support Vector Machine (SVM) that takes deep features and performed activity classification. Dense trajectories were also explored to improve overall network performance [12]. Fan et al. introduced a novel RubiksNet framework for human action recognition in videos. RubiksNet introduced a new learnable shift operation for 3D Spatio-temporal recognition that performs both spatial and temporal shift simultaneously [13]. Zhang et al. proposed a Minimum-Effort-Temporal-Activity-Localization (METAL) framework for untrimmed videos. METAL uses a supervised learning approach to localize unseen activities when only a few examples of a specific class are available. They introduced a Meta Learning-based approach that hierarchically measures similarity matrices among videos and at the same time it localizes and classifies activities [14].

Ullah et al. proposed a long-term activity recognition framework that includes a bi-directional LSTM combined with CNN. For understanding activity intentions they stacked together both forward and backward propagation [10]. Tran et al. developed a 3D CNN model to incorporate temporal information. The convolution operation was performed simultaneously for both spatial and temporal domains. The computational cost for their proposed model was very high despite the use of both spatial and temporal information [15].

Simonyan and Zisserman presented a two streams network. During the 1st stream of network temporal changes information was extracted using a spatial network. The second stream to compute the dense-optical-flow displacement used a temporal network across several frames. At last, for evaluation, they took averages from the results of first and second steams. Most Deep CNN-based approaches used 2-dimensional images and do not consider temporal change information [16]. Wang et al. combined a trajectory feature map and a two-stream network in Trajectory Pooled Deep Convolutional Descriptors (TDD) that combines both shallow local features and Deep Neural Network (DNN) [17]. Zhang et al. presented a combination of seen and unseen data as seen in class regions based Semantic-Similarity-Embedding (SSE). They have also discussed the importance of the label embedding technique as compared to the traditional attributes-based methods. This type of label embedding information is easily available from open text corpora [18].

Du et al. proposed a recurrent Spatio-temporal attention network. They used LSTM for learning spatial-temporal representation of actions. Then they integrated the motion and appurtenance of these LSTMs and then they RSTAN for actor attention. They achieved 95.1% accuracy on UCF101 and 79.9% on HMDB51 datasets [19]. Yang et al. proposed a progressive learning method for spatial-temporal action detection in video data streams. They used spatial refinement and temporal extraction techniques. They used VGG16 on UCF101and 3D convolution layers for temporal modeling and I3D on AVA datasets [5].

Ji et al. proposed a 3-dimensional CNN-based model for recognizing ending human-actions. This 3D CNN model capture features to get motion information from both spatial and temporal dimensions from the several consecutive frames. This approach was relying on the analysis of ending activity sections of humans in video data [20]. Karpathy et al. presented and CNN base framework that learns the motion features based on spatial-temporal information. However, on the UCF-101 dataset, the human activity recognition rate was 63.3%, indicating that their proposed CNN based architecture was not capable of efficiently denotes human activities in visual data streams [21, 22]. Caetano et al. presented another approach, Optical Co-occurrence Matrices, that uses orientation and magnitude attained to collect some of the statistical features of optical flow. The benefit to design Optical Co-occurrence Matric (OFCM) over old approaches was, so that, in the nearby neighborhoods spatial-relationship of field flow can easily be captured.

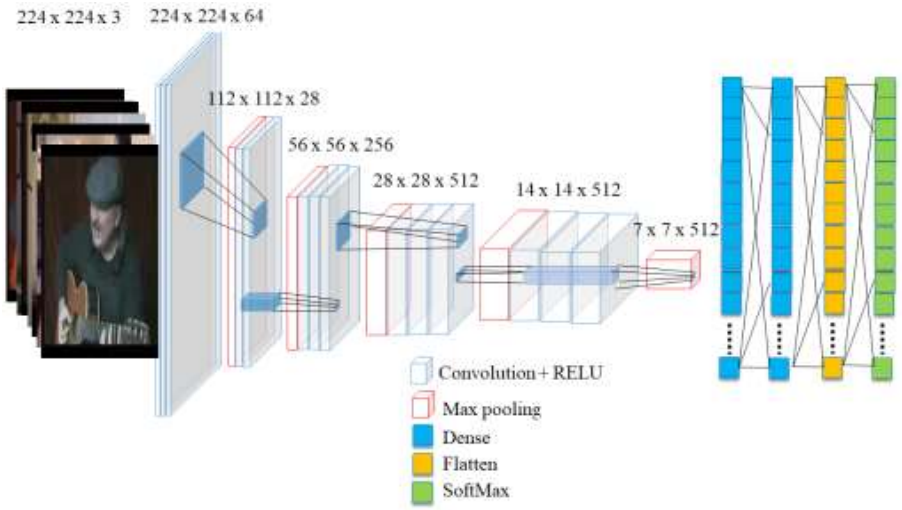

Figure 1: The overall methodology of the activity recognition model

Handcrafted feature extraction involves representing low-level features from the video data, hard engineering, the high difficulty for both abstraction and their classification. Hence, in recent studies automatic feature extraction mechanisms were introduced and adopted by researchers. Instead of handcrafted feature extraction that is complex in the extraction and classification mechanism of features, artificial neural networks-based approaches can directly extract these low-level features from input data. CNN-based approaches automatically learn features in layers iteratively that some early layers capture local-features (edged, pieces of an object) from visual data and some last layers capture global-features (complete object form and structure) representing high-level semantics [9, 23].

Due to the diversity and dynamic nature of video data streams, the use of a previously trained model for action recognition from the newly acquired data is not capable of generating effective results. Lobo et al. presented action recognition on dynamic video data as an optimization problem, which can be solved by using a bio-inspired-heuristic and self-learning optimization technique to handle heterogeneity and high diversity of data streams [24].

III. Proposed Framework

This section describes the mechanism of the proposed system in detail as presented in Figure 1. Video streams consist of a large volume of information hidden in visual data streams that include spatiotemporal variations in motion, colors, texture, and edges. Among many other methods, Deep Learning based methods showed its remarkable effectiveness in image and video content analysis due to its representational abilities. CNN based models need large volumes of data streams and they are highly expensive in terms of computational resources requirements. To overcome these problems, it is better to utilize a pre-trained CNN-based model because its parameters are trained and the model is tuned on some benchmark datasets such as ImageNet. CNN-based approaches automatically learn features in layers iteratively that some early layers capture local-features (edged, pieces of an object) from visual data and some last layers capture global-features (complete object form and structure) representing high-level semantics. The resulting features-vector will be high-dimensional that represent actions from a video sequence in the raw form [9, 25].

A. Dataset

UCF101 [22] is a publicly available benchmark dataset that has 13320 videos from 101 realistic action classes collected in '.avi' format from YouTube. Each class contains 100~300 video samples of an action. The duration of each video ranging from 2~to 7 seconds. UCF0101 is a diverse and challenging dataset where actions were performed by humans with the interaction of different objects, sports items, musical instruments, and interaction with other humans or with the other part of the human body.

B. Data Pre-processing

In data pre-processing, we grouped with those action classes that are similar or they have overlapping actions among them from the UCF101 dataset. Each group contains several classes ranging from 2~to 5. The total number of classes that have overlapping actions among the are six listed in Table I.

After grouping, we extracted frames from each 13320 video and rescaled them to prepare them to input the pre-trained model. For VGG16 all the frames are rescaled into dimension (224,224,3) and for InceptionV3 all the frames are rescaled into dimension (229,229,3).

TABLE I: Overlapping Activities Classes Groups

Group No.	Overlapping Classes
1	ApplyEyeMakeup, ApplyLipstick, ShavingBeard
2	BoxingPunchingBag, BoxingSpeedBag, Punch
3	BalanceBeam, FloorGymnastics, StillRings
4	HighJump, LongJump, FloorGymnastics, JavelinThrow, PoleVault
5	Bodyweight squats, CleanAndJerk, Lunges
6	Shotput, HammerThrow, ThrowDiscus

TABLE II THE EVALUATION RESULTS FOR INVEPTIONV3 AND VGG16 CNN MODELS

	InceptionV3					VGG16				
Group	Accuracy	Loss	Precision	Recall	F1-score	Accuracy	Loss	Precision	Recall	F1-score
1	0.9746	0.0546	0.9760	0.9760	0.9760	0.9975	0.0031	0.9976	0.9976	0.9976
2	0.9976	0.0114	0.9976	0.9976	0.9976	0.9976	0.0074	0.9976	0.9976	0.9976
3	0.9941	0.0151	0.9943	0.9943	0.9943	0.9971	0.0223	0.9972	0.9972	0.9972
4	0.9782	0.0593	0.9807	0.9792	0.9799	0.9813	0.0964	0.9850	0.9747	0.9798
5	0.9971	0.0145	0.9972	0.9972	0.9972	0.9942	0.0184	0.9943	0.9943	0.9943
6	0.9858	0.0330	9866	9866	9866	0.9623	0.1269	0.9767	0.9442	0.9599

C. *Transfer Learning*

Transfer Learning is a technique that allows us to retrain the final layers of a pre-trained model. It drastically reduces dataset requirements and also saves much time for training Deep Neural Network from start. Two of the most famous CNN pre-trained model VGG16 and InceptionV3 are used in this paper. It helps us utilizing these models for our dataset in much less time and with improved accuracy. These pre-trained models are trained on the 'ImageNet' dataset that has 1,000 classes of over one-million images. We utilized the models by adding some layers at the end of the model and removing the last Fully Connected Layers from the original model. Using these pre-trained models in such a way allows us to take advantage of the knowledge that these models learned during their days and weeks of original training and apply that knowledge to our dataset. This setting drastically improves accuracy by training the model with some knowledge instead of starting blank.

D. *VGG-16 CNN model*

A VGG-16 [26] pre-trained CNN model is used in the proposed framework. Pre-trained CNN models learn deep features and are also used for classification. A stack of 'conv' layers was used in architecture. The input for the first layer is of size, (224, 224, 3). Kernel size is (3 x 3) for all convolutional layers in VGG-16 with a stride value of 1. The second convolution layer is also of Kernel size (3 x 3) without any pooling layer in between them. Thus, after the first 2 layers, we have the resulting kernels of size (7 x 7). Three convolutional layers three times are then assembled consecutively with 'max pooing' between them and with a 'relu' activation following them. Three Fully-Connected Layers (FCL) are stacked at end of all other layers. The First 2 FCL has 4096 channels and the last FCL has 1000 classes for the 'ImageNet' pre-trained model.

E. *INCEPTION-V3 CNN model*

InceptionV3 [27] pre-trained model is 42 layers deep. The main focus of this model is to use less computational power. InceptionV3 has certain charsets that are:

- Factorized convolutions: Traditional (7 x 7) convolutions are factorized into 3 (3 x 3) convolutions. It decreases the number of parameters involved in the neural network for the reduction of computational efficiency and to check on the network efficiency.
- Grid size reduction: Perform bottlenecks of computation cost by applying pooling operations. Grid size is compacted to (17 x 17) having 786 filters.
- Smaller convolutions: Instead of using larger convolutions it uses small convolutions that make the training faster.
- Asymmetric convolutions: A good idea is to use a (1 x 3) followed by another (3 x 1) convolution instead of a (3 x 3) convolution.
- Auxiliary classifier: Plays the role of a Regularizer. Small CNN is added among layers during training and the loss computed is then included in the main network loss.

Zero paddings were used for convolutions and Inceptions modules. Then input of the first layer is (229, 229,23). Pooling and padding were added after three consecutive convolutional layers. Three inception modules were stacked after the next three convolutional layers. In the end, pooling was added following by then linear 'logits' and a SoftMax classifier with 1,000 output classes for the 'ImageNet' dataset.

F. *Proposed models*

Taking benefits from Transfer Learning pre-trained VGG16 and InceptionV3 models trained on the 'ImageNet' dataset are utilized. The last Fully Connected Layer is removed from the VGG16 pre-trained model. One Flatten Layer is added after removing FCL layers, Dropout of 0.5, 'relu' activation, and three Dense layers with the SoftMax at the end of the network. InceptionV3 last Dense layer is removed and one Flatten Layer is added, Dropout of 0.5, 'relu' activation, and three Dense layers with the SoftMax at the end of the network. After some minor pre-processing the frames from the video data are passed to both models. Stochastic Gradient Descent (SGD) is used as an optimizer and for loss, Categorical Cross-entropy is used for both models. At last, both VGG16 and InceptionV3 models are trained with added layers at the end of the model.

IV. RESULTS AND DISCUSSION

A total of 20 classes separated into 6 groups shown in Table I are used in training for both VGG16 and InceptionV3 models. The proposed system was implemented using 'Python 3' on Windows 10 environment, Intel(R) Core (TM)

i7-8700 set up with 16 GB RAM, and 8 GB NVIDIA GeForce GTX 1070 GPU. For Deep Learning implementation TensorFlow 2.x was used. The proposed system was evaluated using five different evaluation metrics that are: Accuracy, loss, precision, recall, f-measure score given in Table II.

The system proposed, is evaluated on 20% video samples from the UCF101 dataset. We fed 5 frames-per-second and a batch of 64 frames-at-a-time from activities videos to our proposed system to gain all the benefits of parallelism in GPU.

The complex activities that have overlapping actions in them are hard to distinguish and correctly classify. Due to variation in viewpoint, camera motion, and scale of the actor, the confidence changes over intervals of visual data streams for human action recognition prediction. The transition from one action to another action or abrupt change causes a reduction in confidence score. Iteratively fine-tuning the Transfer Learning-based Human Activity Recognition model (TLHAR) makes it possible to adopt variations in different activities in a non-stationary environment. Our proposed TLHAR models utilize their knowledge from pre-training and show a drastic improvement in the activity recognition task.

V. CONCLUSION

In this paper, we presented an optimized Transfer Learning-based Human Activity Recognition (TLHAR) framework for video data streams in a surveillance environment. For experimentation purposes, we used UCF101, a benchmark video dataset. Pre-training knowledge and semantic features of VGG16 and InceptionV3 CNN models are used for Spatio-temporal activity learning. Experiments show that our proposed system can efficiently learn and classify in the non-stationary heterogeneous nature of video data streams. In the future, multiple actions are performed by multiple actors in a single video stream. Furthermore, Multiview activity recognition is also a good motivation to work on complex datasets.

REFERENCES

[1] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, "From action to activity: sensor-based activity recognition," Neurocomputing, vol. 181, pp. 108-115, 2016.

[2] L. Zhang et al., "ZSTAD: Zero-Shot Temporal Activity Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 879-888.

[3] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer LSTM," IEEE Transactions on Industrial Electronics, vol. 66, no. 12, pp. 9692-9702, 2018.

[4] C. Liang, D. Liu, L. Qi, and L. Guan, "Multi-Modal Human Action Recognition With Sub-Action Exploiting and Class-Privacy Preserved Collaborative Representation Learning," IEEE Access, vol. 8, pp. 39920-39933, 2020.

[5] X. Yang, X. Yang, M.-Y. Liu, F. Xiao, L. S. Davis, and J. Kautz, "Step: Spatio-temporal progressive learning for video action detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 264-272.

[6] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," Future Generation Computer Systems, vol. 96, pp. 386-397, 2019.

[7] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," IEEE Transactions on Multimedia, 2020.

[8] S. Luo, H. Yang, C. Wang, X. Che, and C. Meinel, "Real-time action recognition in surveillance videos using ConvNets," in International Conference on Neural Information Processing, 2016: Springer, pp. 529-537.

[9] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," Neurocomputing, vol. 288, pp. 30-42, 2018.

[10] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," IEEE Access, vol. 6, pp. 1155-1166, 2017.

[11] J. Yu et al., "A discriminative deep model with feature fusion and temporal attention for human action recognition," IEEE Access, vol. 8, pp. 43243-43255, 2020.

[12] S. Nazir, Y. Qian, M. Yousaf, S. A. Velastin Carroza, E. Izquierdo, and E. Vazquez, "Human Action Recognition using Multi-Kernel Learning for Temporal Residual Network," 2019.

[13] L. Fan et al., "RubiksNet: Learnable 3D-Shift for Efficient Video Action Recognition," in Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[14] D. Zhang, X. Dai, and Y.-F. Wang, "METAL: Minimum Effort Temporal Activity Localization in Untrimmed Videos," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3882-3892.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489-4497.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568-576.

[17] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1933-1941.

[18] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4166-4174.

[19] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1347-1360, 2017.

[20] 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.

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732.

[22] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[23] C. Caetano, J. A. dos Santos, and W. R. Schwartz, "Optical Flow Co-occurrence Matrices: A novel spatiotemporal feature descriptor," in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016: IEEE, pp. 1947-1952.

[24] J. L. Lobo, J. Del Ser, E. Villar-Rodriguez, M. N. Bilbao, and S. Salcedo-Sanz, "On the creation of diverse ensembles for nonstationary environments using bio-inspired heuristics," in International Conference on Harmony Search Algorithm, 2017: Springer, pp. 67-77.

[25] K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, and S. W. Baik, "Convolutional neural networks based fire detection in surveillance videos," IEEE Access, vol. 6, pp. 18174-18183, 2018.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.