Neurocomputing 435 (2021) 321–329

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Conflux LSTMs Network: A Novel Approach for Multi-View Action Recognition

Amin Ullah [a], Khan Muhammad [b], Tanveer Hussain [a], Sung Wook Baik [a,*]

[a] Sejong University, Seoul, Republic of Korea
[b] Department of Software, Sejong University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 13 September 2019
Revised 23 November 2019
Accepted 9 December 2019
Available online 10 December 2020
Communicated by Xiaoli Li

Keywords:
Artificial intelligence
Deep learning
Action recognition
Multi-view video analytics
Sequence learning
LSTM
CNN
Multi-view action recognition

ABSTRACT

Multi-view action recognition (MVAR) is an optimal technique to acquire numerous clues from different views data for effective action recognition, however, it is not well explored yet. There exist several challenges to MVAR domain such as divergence in viewpoints, invisible regions, and different scales of appearance in each view require better solutions for real world applications. In this paper, we present a conflux long short-term memory (LSTMs) network to recognize actions from multi-view cameras. The proposed framework has four major steps; 1) frame level feature extraction, 2) its propagation through conflux LSTMs network for view self-reliant patterns learning, 3) view inter-reliant patterns learning and correlation computation, and 4) action classification. First, we extract deep features from a sequence of frames using a pre-trained VGG19 CNN model for each view. Second, we forward the extracted features to conflux LSTMs network to learn the view self-reliant patterns. In the next step, we compute the inter-view correlations using the pairwise dot product from output of the LSTMs network corresponding to different views to learn the view inter-reliant patterns. In the final step, we use flatten layers followed by SoftMax classifier for action recognition. Experimental results over benchmark datasets compared to state-of-the-art report an increase of 3% and 2% on northwestern-UCLA and MCAD datasets, respectively.

1. Introduction

The giant increase of surveillance cameras with variable scope of installation including offices, public places, and roads, is a key source of big video data generation. Exploiting this Big Data for various tasks, such as video retrieval, video summarization [1], violence detection [2], and action recognition [3] are of keen interest for researchers [4]. Human action recognition (HAR) refers to the prediction of the action status of human in a given video and is the center of concentration for many computer vision scientists, due to its inclusive range of applications. Its applications consist of surveillance, security and law enforcement, videos retrieval, video summarization, and human–computer interactions [5]. Besides the wide range of applications, it has several challenges, such as similar visual contents, viewpoint changes, variable targets, poses and scales of the action performers, and different illumination conditions.

HAR domain is brodly divided into two broad categories based on the number of cameras that capture the motion of the target. The first is single-view action recognition, which has one camera for moving targets and the second category is multi-view action recognition (MVAR) comprising multiple cameras focused on the target. There are a lot of approaches for single-view action recognition and deep learning based methods are particularly abundant [6]. In contrast to single-view action recognition, MVAR is more challenging, because the variation in the features from different viewpoints and invisible regions of appearance in each view yields poor classification results [6–8]. Many scientists proposed valuable researches contributions in this domain with different feature engineering approaches and effective classifiers. For instance, to improve the performance of self-similarity based methods, Yan et al. [9] proposed a multitask learning framework for MVAR. Their framework has a special mechanism to share the self-similarity matric among different views. The authors tested their method over multi-view RGB and RGBD datasets to show better results compared to the state-of-the-art methods. Baradel et al. [10] introduced a visual attention module, which can learn to predict glimpse sequences, that are further processed to achieve final and distributed tracking/recognition results. They used an RGBD

* Corresponding author at: Daeyang AI center, Room# 411, 209 Neungdong-ro, Gwangjin-gu, Seoul, Republic of Korea.
E-mail address: sbaik@sejong.ac.kr (S.W. Baik).

https://doi.org/10.1016/j.neucom.2019.12.151
0925-2312/

and a northwestern-UCLA multi-view action 3D datasets to prove the validity of their system. Classifying human actions from distributed views is difficult due to huge appearance in variations of different views. Liu et al. [11] considered the learning of discriminant view-invariant representations as a key to this problem, which generalizes well over different views. They solved it through the learning of view-invariant representations hierarchically using their proposed concept of joint sparse representation and distribution adaptation. Their experiments over four multi-view datasets outperformed the employed MVAR approaches. A novel concept of dividing and aggregating a network (DA-Net) is presented in [12] for MVAR that learns the independent view representations to share among all the views at lower layers and single view-specific representation for each view at higher layers. The view-specific action classifiers based on each view representations are used to predict how likely each video belongs to a respective view. Finally, these predicted probabilities of the view-specific action classifiers are used as weight and fused for final decision. Similarly, many other approaches are presented for cross-view action recognition based on various flavors of CNNs. For instance, Xiao et al. [13] incorporated feature learning using their novel CNN model on multi-view dynamic images. The dynamic images captured from different cameras are processed via same convolutional layers, but their response is different to the fully connected layers. The authors performed experiments on three challenging datasets to prove its validity and better performance compared to the available techniques. An electromyography based MVAR framework along with electromyography-vision action dataset is presented in a recent paper [14]. A multi-view benchmark dataset for HAR along with an evaluation of the different learning problems is presented in [15]. In this paper, the authors avoided the idea of training model on the source and utilizing it for targets with the intuition that the meaning of each feature dimension is yielding very different results, which were inconvenient. Finally, they used multi-task learning to discover the common knowledge among the underlying views for action recognition.

A keen observation and challenges of the current literature revealed several limitations of the employed approaches that are aimed to be addressed in the current research proposal. The first challenge is the computation of the stable features across multi-views, for which researchers have proposed attention modules, joint sparse representation with distribution adaptation, and multi-view dynamic images. However, these techniques are very sensitive to the large-scale and viewpoint related changes in multi-view data. Furthermore, all these techniques lack the priority of the appearance and motion information carried out in all the views at once that is very important and necessary for the decision based on multi-view data. Similarly, MVAR literature is deprived from techniques that can extract effective temporal information and also retain the time complexity of the trained model as low as possible.

To dominate the MVAR state-of-the-art results and contribute to the action recognition literature, we introduced a conflux LSTMs network based framework. Our proposed method is applicable for RGB data, which has several advantages over other approaches, such as skeleton data and many others. The major benefits for using RGB data is its adaptability and availability, which is practical and easily implementable. The main idea of our framework is the usage of view self-reliant and view inter-reliant processing with parallel LSTMs for each view as visualized in Fig. 1. The key contributions of this work are listed as follows:

1. The frame level representation is very crucial in video analytics. For sequence learning in action recognition, it is important to capture the tiny local changes in continuous frames. For this purpose, we investigated the intermediate layers of a pre-trained VGG19 CNN model. After an in-depth analysis of all the layers, we chose the Conv5_4 layer for features extraction, which can effectively capture the local representation in an image.
2. We proposed a conflux structure of LSTMs network, which has separate LSTM for each view and processes the sequential features obtained from the consecutive frames. This structure allows our network to learn the view self-reliant sequential patterns by processing single view data.
3. We exploited the view inter-reliant processing layers, which find the inter-view correlations by fusing the self-reliant sequences patterns via their pairwise dot product. The view inter-reliant sequences are further processed by the fully connected layers for effective action recognition.

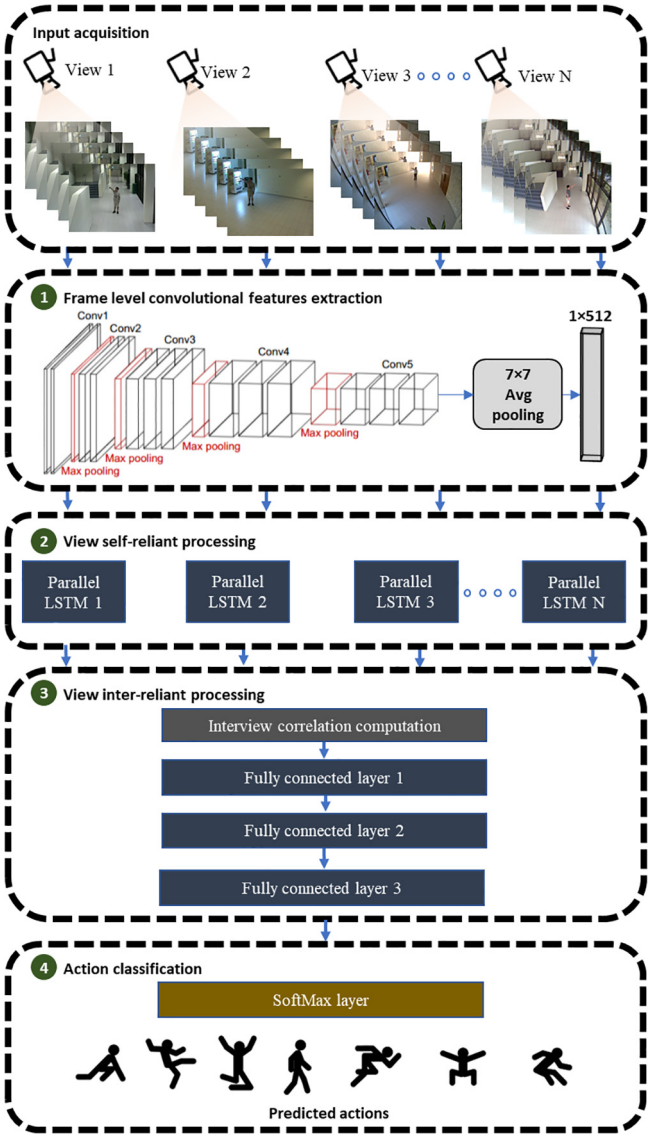

Fig. 1. The proposed MVAR framework using a conflux LSTMs network. Firstly, a sequence of frames from each view is passed to the CNN model for frame level features extraction. Secondly, the extracted features from each view are passed to a self-reliant LSTMs network for sequence learning. Next, the outputs of all the LSTMs are combined via view inter-reliant layers. Finally, the actions are recognized using a SoftMax classifier.

The rest of the paper is organized as follows. The sequential features extraction and the structure of the proposed conflux LSTMs network are discussed in Section 2. The experimental evaluation

and discussion about results are given in Section 3. Section 4 summarizes the key findings of this article and recommends the future research directions.

2. Proposed methodology

In this section, we discussed the implementation procedure of the proposed conflux LSTMs framework for MVAR. First, we acquire video data from multi-view cameras. Second, we forward the sequences for frames from each view to the pre-trained VGG19 CNN model to extract frame level features using our suggested procedure from VGG19 intermediate convolutional layer. Next, the view self-reliant LSTM network processes these features and forward propagates it to the view inter-reliant network for the final action classification. All the steps are subsequently discussed in detail, the overall framework is visualized in the Fig. 1, and mathematically presented in Algorithm 1.

2.1. Convolutional features extraction for sequence representation

CNNs are considered among the principal architectures to represent visual data effectively as compared to hand crafted features representation techniques. The recent achievements of CNNs for large-scale image classification [16], facial recognition [17], and image retrieval [18] have encouraged the computer vision community to exploit it for different domains of data in order to effectively learn complex patterns. The hierarchical structure and its learned kernels give them a powerful means to extract higher level of representations from visual data. The hierarchy of layers allows the neurons of each layer to learn the distinctive features. For instance, the layers at deeper position in the CNN architecture learn more complex and global features, which are known as fully connected layers and are mainly utilized for recognition tasks. Likewise, the initial layers are more sensitive to the local features because the receptive fields of convolutional kernels cover only a tiny portion of an image, which are essentially utilized for dynamic nature features representation tasks. In terms of video data, the global information in continuous frames change very slowly, so the fully connected layer gives almost the same features for the successive video frames as it extracts global higher-level abstractions. On the other hand, the consecutive frames have a lot of local motion dynamics, which can be easily captured via convolutional features, because its kernels convolve a small respective field of image. Therefore, we have utilized the convolutional features for frame level representation in the proposed MVAR framework.

The convolutional layers are considered as the backbone of deep CNN models, because they learn sharp and tiny 2D patterns in an image via the different sizes of kernels. We have encountered many recent studies, which utilized the convolutional layers in image representation for various computer vision applications that include image retrieval [19], fire detection and localization [20], shot segmentation [21], and action recognition [3]. For instance, Jamil et al. [19] presented a framework for object-oriented features selection from convolutional maps of a pre-trained CNN model for image retrieval. In addition, Zhou et al. [22] presented a framework for object detection, which showed that the combination of different feature maps can be utilized for object detection. Similarly, Khan et al. [20] utilized convolutional feature maps for fire detection and localization. These researchers have claimed that the convolutional features are dominant over the fully connected layer features and it is also evident from fact that famous ResNet and MobileNet CNN models avoided fully connected layers in their architectures. In this study, we investigated the convolutional layer of a pre-trained VGG19 [23] CNN model for frame level representation in MVAR task. The VGG19 has a five-level hierarchy of convolutional blocks, each block consists of multiple convolutional and pooling layers, which aim at achieving effective patterns learning from image data. We utilized the final layer of block five (Conv5_4) of the VGG19 CNN model that covers the largest receptive field of an image and assembles feature maps to represent the abstraction of the whole image. The convolutional features can be extracted for feature maps as formula given in Equation (1).

$$C_F(K) = \frac{1}{(w.h)} \sum_{i=1}^{w} \sum_{j=1}^{h} FM^K(i,j) \tag{1}$$

The final layer of block five has $7 \times 7 \times 512$ feature maps, which are fed to Eq. (1), where $C_F(K)$ is the convolutional feature vector, K is the index of feature maps FM, w is weight, and h is the height of FM. Also, the mechanism is given in Fig. 2. We acquired a 512-dimensional feature vector for a single video frame. In the proposed framework, we processed a sequence of 15 frames parallel from each view for the MVAR. Therefore, 7680-dimensional feature vector is fed to our conflux LSTMs network for sequence learning from each view.

2.2. View self-reliant network

In this section, we explain the structure of the proposed conflux LSTMs network and its effectiveness in terms of sequence learning from multi-view data for action recognition. As in a multi-view cameras scenario, the data is captured from different angles, so each view has different visual information. For instance, if a person is performing an action in front of different cameras, so the camera with his frontal orientation will have the most important visual information [1]. Therefore, keeping this fact in our minds and different from state-of-the-art, we first individually processed each view data to learn the action patterns that we termed as view self-reliant network, and then finally combined all the clues collected from each view for action recognition. The LSTMs are more powerful than the feed-forward neural networks for learning parametric and structural patterns to build a sequential model. It can process the arbitrary sequences of the time series input and exhibits a dynamic behavior by committing previous information to its memory and using it for final output based on the maintained history. Moreover, it has the ability to capture the intricate interactions between neighborhood space through its hierarchical hidden layers structure [24]. Furthermore, the recent studies on LSTMs have proven that its multilayer and bidirectional structures are more effective in terms of long terms sequence learning prob-

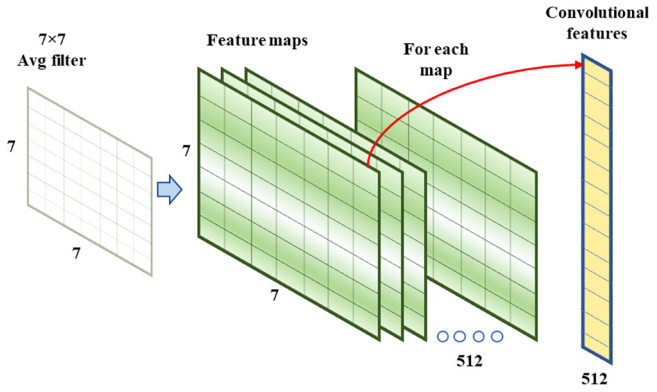

Fig. 2. The features extraction mechanism, where a 7×7 average filter is applied on the feature maps of a VGG19 CNN model to acquire the convolutional feature vector for the frame level representation. There are 512 feature maps in the Conv5_4 layer, so this mechanism generates a single representative value from each map and generates a final convolutional featurevector.

lem. However, this type of structure increases the time complexity of networks and are not applicable for real-time applications. Therefore, keeping all these facts in our mind, we proposed a multilayer LSTM structure for each view in self-reliant network. The configuration details of different layers of our conflux LSTMs network are given in Table 1.

Our network takes visual features extracted from consecutive frames $\{f_1, f_2, f_3, f_4, \cdots, f_n\}$using a CNN model for each view $\{V_1, V_2, V_3, V_4, \cdots, V_n\}$ and assumes that all cameras are synced while capturing images. The network has $\{L_1, L_2, L_3, L_4, \cdots, L_n\}$ multilayer LSTMs, where $L_i \in V_i$ and results in sequence to sequence output. The multilayer LSTM structure for each view learns multi-view independent sequential patterns. The detailed explanation about the structure of LSTM is out of scope of this paper, but it is important to discuss the structure of our multilayer LSTM. For each $L_i \in L$, we have three layers stacked LSTM $\{l_1, l_2, l_3\} \in L_i$. The l_j contains 256 memory cells, takes a 512-dimensional feature vector, and outputs a sequence equal to $\left(\frac{length\ of\ memory\ cell}{2}\right)$. The output sequence from l_j is further processed by the view inter-reliant network for MVAR.

2.3. View inter-reliant network

After the supplementary patterns learning through the view self-reliant LSTMs, it is important to capture the higher-level dependencies in multi-view sequences in conjunction. Due to the overlapping locations and the fields of view of different cameras, there is a huge amount of correlations in the multi-view data. Therefore, exploiting these correlations play a significant role and work as strong recognition clues from each view. Furthermore, these correlations and inter-dependencies between them need to be accurately modeled for effective MVAR. Our view self-reliant network outputs representative sequences $\{S_1, S_2, S_3, \cdots, S_n\}$ for each of the view data, which contain strong correlations that are sought out by our view inter-reliant network in conflux LSTMs settings. The inter-view correlation between the two feature vectors has been investigated by plenty of researchers. For instance, Panda and Roy [25] utilized multi-view embeddings to capture the multi-view correlations using sparse coefficients. Similarly, Hussain et al. [1] presented a novel idea of lookup table, which stores frames from multiple views in a synchronous manner, which thereby helps their system to avoid extra processing for correlations calculation. A well-known FlowNet [26] CNN model utilized multiplicative patch comparisons for correlations analysis in optical flow generation between two consecutive feature maps. Inspired from the idea of FlowNet, we also utilized pairwise dot product of features maps for correlations computations between sequential features of multiple views.

Table 1
The configuration of our conflux LSTMs network, which summarizes the input and output dimensions for each view LSTM and the number of trainable parameters in the network.

Layer	Dimensions	No. of parameters
Input	$\|\| 15 \quad 512 \|\| \times 3$	–
LSTM (View1)	$\left\|\left\| \begin{matrix} 512 & 256 \\ 256 & 128 \end{matrix} \right\|\right\| \times 3$	2,286,276
LSTM (View2)	$\left\|\left\| \begin{matrix} 512 & 256 \\ 256 & 128 \end{matrix} \right\|\right\| \times 3$	2,286,276
LSTM (View3)	$\left\|\left\| \begin{matrix} 512 & 256 \\ 256 & 128 \end{matrix} \right\|\right\| \times 3$	2,286,276
Correlations	$\|\| 128 \quad 128 \quad 128 \|\|$	–
FC 1	$\|\| 1 \times 128 \|\|$	16,384
FC 2	$\|\| 1 \times 64 \|\|$	4,096
FC 3	$\|\| 1 \times 18 \|\|$	324
SoftMax	$\|\|$ ***no of classes*** $\|\|$	**6,817,220**

In our conflux network, we employed a correlation layer that performed the pairwise dot product between the representative sequences of multi-view data. In this layer, the view self-reliant sequential features are convolved with the features of other sequences instead of convolving it with kernels, as in typical convolutional layer of NNs. For instance, we have n dimensions multi-view sequences $\{S_1, S_2, S_3, \cdots, S_n\}$, for which our correlation layer compares a feature point of $S_1$and S_2. Let us consider a 'c' correlation comparison of feature point $S_i^{V_1}$ with n multi-view sequences as

$$c(S_1, S_2, \cdots, S_n) = \langle (S_i^{V_1}), (S_i^{V_2}), \cdots (S_i^{V_n}) + o \rangle \tag{2}$$

where S represents the sequence from each view, 'i' is the index of feature point to be compared, and 'o' is the bias unit. This layer outputs one dimensional representation for all the views, which is further processed via fully connected layers for MVAR task. We have used the configuration given in Table 1 for three views data, which represents separate LSTM network for each view.

Algorithm 1: Conflux LSTMs Network

Input: Multi-view video streams $\{V_1, V_2, V_3, V_4, ..., V_n\}$
Output: Predicted action class along with the probability score
Preparation:
1. Acquire synchronized multi-view frames
2. Load pretrained VGG19 CNN model M_1
3. Initialize trained Conflux LSTMs network M_2

Steps:
while (video frames $(\{V_1, V_2, V_3, V_4, ..., V_n\})$)
1. Read frames $\leftarrow \{f_i \in V_j\}s$
2. Forward $f_i \in V_n$ frames to M_1
3. $FM \leftarrow M_1$
4. Apply 7×7 average pooling to $FM^K$for frame level visual features extraction using Eq. (1).
5. Repeat step 2, 3, and 4 for sequence of frames of V_n.***note:** sequence length in our experiments is 15
6. Combine frame level sequential features F_v from $\{V_1, V_2, V_3, V_4, ..., V_n\}$
7. Labeled action class $\leftarrow$ Forward propagate F_v to M_2
8. Show predicted action along with probability score

end while

The LSTMs structure is deeply discussed in the view self-reliant network. Our view inter-reliant network depends on the output of correlation and the fully connected layers. We stacked three fully connected layers with 128, 64, and 18 dimensions after the correlation layer, which captured the higher-level abstraction from multi-view data for combined action recognition. The conflux LSTMs network has total 6.8 million parameters for three views data. The higher is the number of views in the conflux network, higher will be the network size and computational complexity. Also, the accuracy of the network varies by changing the number of views and from the experiments we observed that the effectiveness of the network depends on the overlapping invisible regions and the different scales of human appearance from the camera.

3. Experimental results and discussion

In this section, we discussed different experiments performed using MCAD [27] and northwestern-UCLA [28] MVAR benchmark datasets for the evaluation of our conflux LSTMs network. We investigated our method for overall recognition accuracy, the confusion matrix, the class-wise performance, the trained deep learning model size, and the time complexity of our network in different

multi-view scenarios. We also compared our results with state-of-the-art and discussed some scientific reasons for the dominance of our network. The experiments of the proposed conflux network were performed in Python 3.5 with the deep learning library Tensorflow-1.12, installed over Ubuntu-16.04. The hardware equipped for the experiments contained a CoreTMi5-6600 processor with 16 GB RAM and supplied with the support of a dedicated 12 GB GeForce-Titan-X GPU.

3.1. Multi-camera action dataset (MCAD)

The mainstream action recognition datasets are created for different purposes based on the applications and the purpose, such as sports and entertainment actions, consumer generated actions, and surveillance datasets. However, the multi-camera action is recognized from surveillance data that is captured through multiple cameras in a CCTV environment. In an MCAD dataset [27], the videos are recorded using five cameras that are installed with different angles to capture the overlapping areas. It has two types of camera settings, which include the static and the pan-tilt-zoom cameras. There are three static cameras with fisheye, which include Cam04, Cam05 and Cam06, and two pan-tilt-zoom cameras, which include PTZ04 and PTZ06. The resolutions of the static and the PTZ cameras are 1280×960 and 704×576, respectively. Moreover, the contrasting effect is added as day and night. A total of 18 actions are recorded by 20 recruited individuals, where each individual repeats the action 8 times, such as 4 times during the day and 4 times during the evening. It mainly has two types of actions, such as single-person action and person-to-object action. It is a more challenging dataset because out of the five cameras, the two moving and zooming cameras data increase the viewpoint changes and create scaling issues in sample.

3.2. Northwestern-UCLA multi-view action 3D dataset

This dataset [28] is recorded with three simultaneous Kinect cameras by the University of California and Northwest University in Los Angles. The settings used for its recording are same as the Multiview 3D Event dataset, but multiple locations are added. It comprises of RGB, depth and human skeleton data of 10 action categories performed by 10 individuals, captured from different viewpoints. In this paper, we utilized only RGB data of the dataset for MVAR. The action categories include pick up with one hand, pick up with two hands, dropping the trash, walking around, doffing, throwing, and carrying. The visual contents of all categories are very similar and the data is recorded in same background, which make it very challenging. We observed that the only discriminative sequential features, which can be helpful in a challenging environment is the motion of human body parts that are effectively captured in the proposed framework using the local convolutional features.

3.3. F. Training process and parameters selection for conflux LSTMs network

The proposed conflux LSTMs network acquire features from the pretrained CNN model. We investigated features extracted from the fully connected layer and the intermediate convolutional features of CNN. The sequence length for the conflux LSTMs is 15 consecutive frames that is selected after observing its effectiveness and the time complexity analysis of the 30, 25, and 15 frames sequences. Considering 15 frames per sequence using the fully connected layer deep features, our network takes 1000-dimensional features vector at a time step 't' of conflux network and using the convolutional features, it takes 512-dimensional input. The experiments are performed using LSTM with 512 and 256 memory cell sizes, where the optimal selected size is 256 in order to have a lower computational complexity, because the conflux LSTMs network has three-layered stacked LSTMs for each view, and 512 cell size increased the processing time exponentially. The learning rate is a very important factor, which we initialized from 0.01 and after 250 iterations, we reduced it to 0.001, and the stochastic optimization function was used for cost minimization. The datasets are formed as 60% for training and 20% for each validation and testing set of the conflux LSTMs network. The performance of our network during training process is illustrated in Fig. 3. It can be observed from the Fig. 3 that the convolutional features of Covn5_4 have effective results as compared to the deep features of FC8 layer. The deep features are an abstract representation of the frame that cannot capture the slight changes between consecutive frames. On the other hand, the convolutional features are very effective in terms of local changes representation, so its validation accuracies are much higher compared to deep features.

3.4. Closed set, open set, and class-wise evaluation

The closed set and the open set are very useful evaluation metrics to know the robustness of a trained AI model. Nevertheless, the mains stream AI methods only follow the closed set evaluation where the test samples are separated from the training data, even though the test samples are collected from the same environmental settings as the training data. On the other hand, in the open set evaluation the test set is totally different from the training set, and it is collected from different scenarios. In this study, we have assessed our conflux LSTMs network through both closed and open sets. For the closed set, testing samples are separated from the same multi-views data as training, but we trained our model on one view and tested using another view data for the open set. For instance, our conflux LSTMs take at least two views input data, so for closed set, we inputted three views data for the training and separated 20% for testing from the same views. However, for open set we trained on view 1 and view 2 and tested on view 3 and view 4. The confusion metrics for closed set are given in Fig. 4 and the open set are illustrated in Fig. 5. It can be seen from Fig. 4 that for the closed set the true positive intensities are higher for almost all categories of both datasets. However, in Fig. 5(a), the open set shows that the results are reduced a little bit, because the model encountered a completely novel type of data, which was not used during training and the model is confused in the walk around and carry classes. We got a 90.1% overall accuracy for the closed set and 88.9% for the open set on northwestern-UCLA. Similarly, for on the MCAD dataset, we achieved 80.3% accuracy for the closed set and 86.9% accuracy for open set. The class-wise performance by our conflux LSTMs network for northwestern-UCLA dataset is given in Fig. 6(a) and for the MCAD dataset is illustrated in Fig. 6(b). It can be seen from Fig. 6(a) that the bars of all the classes give a better performance and are mostly higher than 70%. The stand up, donning, and doffing classes have all crossed 90%. However, the sit-down class only reached 55% and the reason is very clear from Fig. 4(a) where in confusion matrix, the sit-down class is confused with the standup class.

3.5. Comparison with the state-of-the-art

The proposed conflux LSTMs network is extensively compared with the state-of-the-art methods via the results of different views and the overall accuracy. Table 2 illustrates the comparison with the depth, the pose, and the RGB based methods for the northwestern-UCLA multi-view action dataset. This dataset has only three views, for which the researchers formed an experimental setup, where they trained a model with V_1 and V_2 and tested it with the V_3 data. However, the proposed network processed at

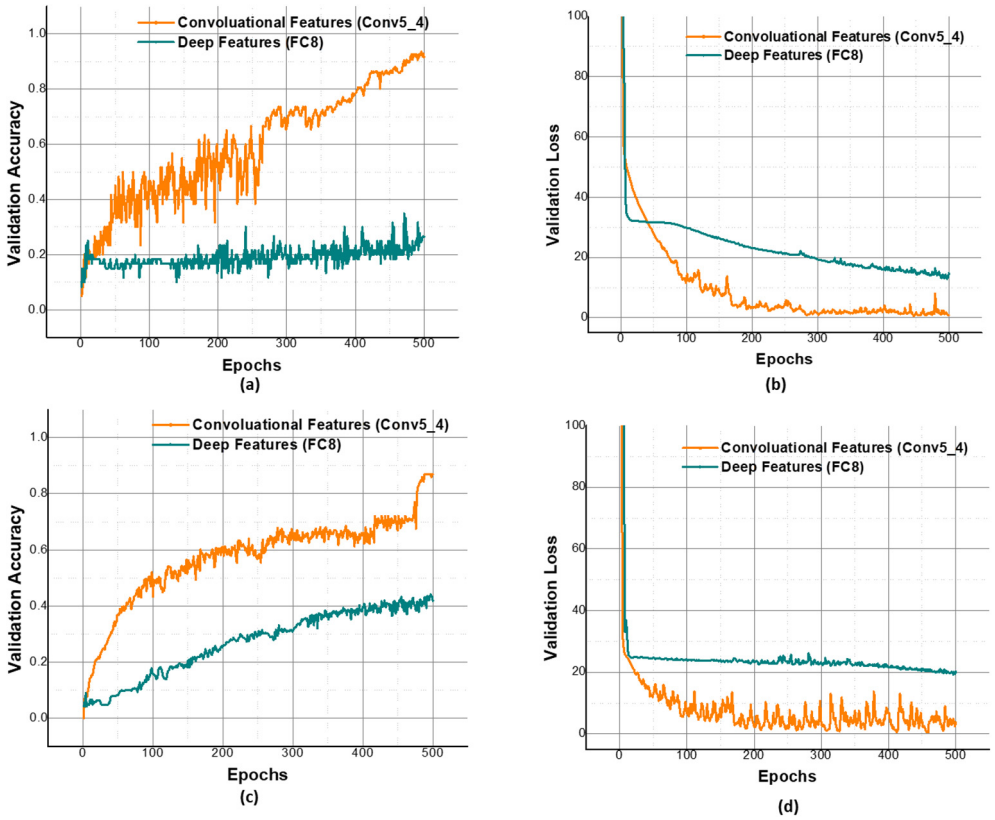

Fig. 3. The validation accuracy and loss achieved at each epoch during the training process of the conflux LSTMs network for (a, b) the MCAD and (c, d) Northwestern-UCLA Multi-view actions datasets.

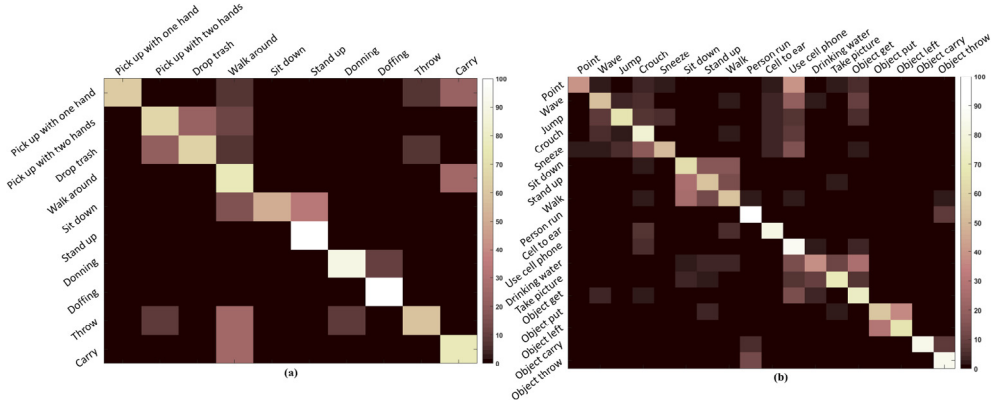

Fig. 4. The confusion matrix for the closed test set of (a) the northwestern-UCLA dataset and (b) MCAD dataset. The bar line displays the accuracy range from 0 to 100 where the classes that achieved a brighter color on its diagonal has better results, and the ones that are closer to a dark color are confused with other classes.

least two views and we cannot forward propagate single view to the network, so we therefore performed the training with V_1 and V_2 and testing with V_2 and V_3. It can be seen from Table 2 that each training and testing gave outputs with different accuracies. In the depth based methods, the 3D viewpoints [31] achieved 91.9% accuracy for setting one, but it got 75.2%, 71.9%, and 79.7% accuracies for setting two, three, and average performance, respectively. The pose based methods performed their experiments using setting one, where the temporal sliding LSTM [34] obtained highest accuracy of 89.2% and a view invariant HAR [33] and a Hierarchical RNN [32] achieved 86.1% and 78.5% accuracies, respectively. The RGB based methods achieved varying accuracies for each view settings where the glimpse clouds [10] obtained 90.1% highest accuracy with setting one and the proposed conflux LSTMs network obtained highest accuracies of 92.5% and 88.6% for settings two and three, respectively. The proposed method also reached the high average performance for all the settings reaching 88.9%. The proposed network has good performance for all the settings, because multi-view data is processed parallelly where it first gets the view self-reliant features and then view inter-reliant features which help our model to learn the features of all the views scenarios for effective MVAR. The comparison using the overall recognition accuracy for northwestern-UCLA and MCAD datasets is given in Tables 3 and 4. The proposed conflux LSTMs network has out-

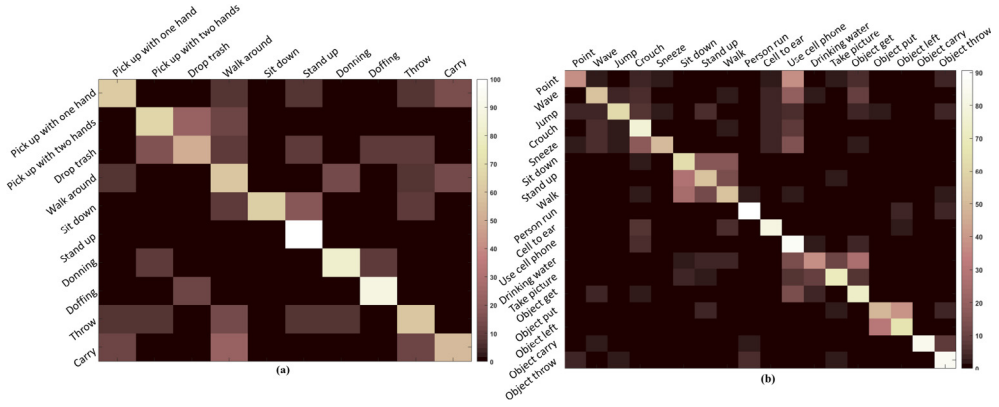

Fig. 5. The confusion matrix for open test set of (a) the northwestern-UCLA dataset and (b) the MCAD dataset.

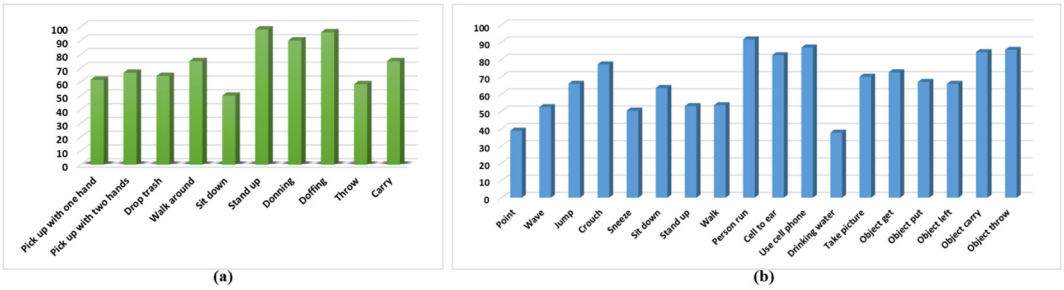

Fig. 6. Class-wise performance of the proposed conflux LSTMs network on (a) the northwestern-UCLA dataset and (b) the MCAD dataset.

Table 2
Comparison of the proposed conflux LSTMs on the northwestern-UCLA multi-view action dataset via different view (V) settings with depth, pose, and RGB based methods, respectively.

Data	Methods	Train V_1 & V_2	Test V_3	Train V_1 & V_3	Test V_2	Train V_2 & V_3	Test V_1	Average
Depth	Virtual views [29]	58.5		55.2		39.3		51.0
	Virtual path [30]	60.6		55.8		39.5		52.0
	3D viewpoints [31]	91.9		75.2		71.9		79.7
Pose	Hierarchical RNN [32]	78.5		–		–		–
	View invariant HAR [33]	86.1		–s		–		–
	Temporal sliding LSTM [34]	89.2		–		–		–
RGB	3D pose motion [35]	68.6		68.3		52.1		63.0
	Knowledge transfer model [36]	75.8		73.3		59.1		69.4
	Glimpse global model [10]	85.6		84.7		79.2		83.2
	Glimpse clouds [10]	90.1		89.5		83.4		87.6
	Conflux LSTMs network	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	
		V_1 & V_2	**V_2 & V_3**	**V_1 & V_3**	**V_2 & V_3**	**V_2 & V_3**	**V_1 &V_3**	
		85.7		**92.5**		**88.6**		**88.9**

Table 3
Comparison with state-of-the-art methods using the overall recognition accuracy of the northwestern-UCLA multiview-3D dataset.

Method	Accuracy (%)
MST-AOG w/o Low-S [28]	65.3
MST-AOG w Low-S [28]	73.3
HOPC [37]	80.0
Multi-view dynamic images + CNN [13]	84.2
Conflux LSTMs network	**88.9**

Table 4
Comparison with state-of-the-art methods using the overall recognition accuracy of the MCAD dataset.

Method	Accuracy (%)
IDT [38]	84.2
Covariance matrices [39]	64.3
STIP [27]	81.7
Cuboids [27]	56.8
Conflux LSTMs network	**86.9**

performed state-of-the-art on both datasets. It improved 5% on northwestern-UCLA dataset, reaching 90.1% from 84.2%, which was previously achieved by [13]. Similarly, on MCAD dataset our model improved 3% accuracy from the level previously achieved by IDT [38]. The results above discussed indicate a better performance and robustness of our conflux LSTMs network in any sort of multi-view cameras settings.

4. Conclusion and future work

In this paper, we introduced a novel concept of conflux LSTMs for MVAR. It is a challenging area of research with several applications that range from daily life surveillance to the monitoring of cities. Our framework incorporates several LSTMs in a network to recognize action from multi-view cameras. There are four major steps in our framework, which include (1) preprocessing, (2) a conflux LSTSMs network for view self-reliant patterns learning, (3) inter-view correlation computation for view inter-reliant patterns learning, and (4) action classification. In the first step, we passed a sequence of frames to the CNN model and extracted a 1×512 feature vector from an intermediate convolutional layer. The feature extraction from the convolutional layer is more advantageous and robust for our problem due to the slight changes in the frame-level features from the fully connected layers. In the second step, these features are inputted to the LSTMs network to compute the view self-reliant patterns. The inter-view correlations among the different views are very important for MVAR. Therefore, we computed the inter-view correlation in the third step by taking the pairwise dot product from the output of the LSTMs network respective to each view. Finally, we classified the underlying action through a SoftMax classifier by passing the features of the flattened layers of our conflux LSTMs network. The experimental results compared to the recent state-of-the-art methods are dominating, proving that our framework can be helpful in many real-time applications.

The multi-view data is of very high dimensions and processing it via deep architecture leads to high computations. In the future, we want to replace our features extraction model with an alternative light-weight model to have faster computations and with intention of presenting a more optimal conflux structure for the MVAR. We also want to try embedded programming [40] in order to transform our framework into a resource-constrained device that can be fitted anywhere for better MVAR. Furthermore, we will also consider other sensors data [41–43] along with vision data by developing some fusion mechanism for effective action recognition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2019R1A2B5B01070067).

References

[1] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S.W. Baik, V.H.C.d. Albuquerque, Cloud-assisted multi-view video summarization using CNN and Bi-Directional LSTM, IEEE Trans. Ind. Inf. 16 (1) (2020) 77–86, https://doi.org/10.1109/TII.2019.2929228.

[2] F.U.M. Ullah, A. Ullah, K. Muhammad, I.U. Haq, S.W.J.S. Baik, Violence detection using spatiotemporal features with 3D convolutional neural network, Sensors 19 (11) (2019) 2472.

[3] A. Ullah, K. Muhammad, I.U. Haq, S.W. Baik, Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments, Future Generation Comput. Syst. 96 (2019) 386–397.

[4] M. Majd, R. Safabakhsh, Correlational convolutional LSTM for human action recognition, Neurocomputing 396 (2019) 224–229, https://doi.org/10.1016/j.neucom.2018.10.095.

[5] Z. Chen, L. Zhang, C. Jiang, Z. Cao, W. Cui, WiFi CSI based passive human activity recognition using attention based BLSTM, IEEE Trans. Mob. Comput. 18 (11) (2019) 2714–2724, https://doi.org/10.1109/TMC.2018.2878233.

[6] W. Sui, X. Wu, Y. Feng, Y. Jia, Heterogeneous discriminant analysis for cross-view action recognition, Neurocomputing 191 (2016) 286–295.

[7] J. Zheng, Z. Jiang, R. Chellappa, Cross-view action recognition via transferable dictionary learning, IEEE Trans. Image Process. 25 (6) (2016) 2542–2556.

[8] Z. Chen, L. Zhang, Z. Cao, J. Guo, Distilling the knowledge from handcrafted features for human activity recognition, IEEE Trans. Ind. Inf. 14 (10) (2018) 4334–4342.

[9] Y. Yan, E. Ricci, R. Subramanian, G. Liu, N. Sebe, Multitask linear discriminant analysis for view invariant action recognition, IEEE Trans. Image Process. 23 (12) (2014) 5599–5611.

[10] F. Baradel, C. Wolf, J. Mille, G.W. Taylor, Glimpse clouds: human activity recognition from unstructured feature points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 469–478.

[11] Y. Liu, Z. Lu, J. Li, T. Yang, Hierarchically learned view-invariant representations for cross-view action recognition, IEEE Trans. Circuits Syst. Video Technol. (2018).

[12] D. Wang, W. Ouyang, W. Li, D. Xu, Dividing and aggregating network for multi-view action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 451–467.

[13] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J.T. Zhou, X. Bai, Action recognition for depth video using multi-view dynamic images, Inf. Sci. 480 (2019) 287–304.

[14] L. Wang, B. Sun, J. Robinson, T. Jing, and Y. Fu, "EV-Action: Electromyography-Vision Multi-Modal Action Dataset," arXiv preprint arXiv:1904.12602, 2019.

[15] Z. Gao, T.-T. Han, H. Zhang, Y.-B. Xue, G.-P. Xu, MMA: a multi-view and multi-modality benchmark dataset for human action recognition, Multimedia Tools Appl. 77 (22) (2018) 29383–29404.

[16] O. Russakovsky et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.

[17] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[18] I. Mehmood, Amin Ullah, Khan Muhammad, Der-Jiunn Deng, Weizhi Meng, Fadi Al-Turjman, Muhammad Sajjad, Victor Hugo c. De Albuquerque, Efficient image recognition and retrieval on IoT-assisted energy-constrained platforms from big data repositories, IEEE Internet Things J. 6 (6) (2019) 9246–9255.

[19] J. Ahmad, K. Muhammad, S. Bakshi, S.W. Baik, Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets, Future Generation Comput. Syst. 81 (2018) 314–330.

[20] K. Muhammad, J. Ahmad, S.W. Baik, Early fire detection using convolutional neural networks during surveillance for effective disaster management, Neurocomputing 288 (2018) 30–42.

[21] I.U. Haq, K. Muhammad, A. Ullah, S.W. Baik, DeepStar: detecting starring characters in movies, IEEE Access 7 (2019) 9265–9272.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," arXiv preprint arXiv:1412.6856, 2014.

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

[24] E. Tsironi, P. Barros, C. Weber, S. Wermter, An analysis of Convolutional Long Short-Term Memory Recurrent Neural Networks for gesture recognition, Neurocomputing 268 (2017) 76–86.

[25] R. Panda, A.K. Roy-Chowdhury, Multi-view surveillance video summarization via joint embedding and sparse optimization, IEEE Trans. Multimedia 19 (9) (2017) 2010–2021.

[26] A. Dosovitskiy et al., Flownet: learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.

[27] W. Li, Y. Wong, A.-A. Liu, Y. Li, Y.-T. Su, M. Kankanhalli, Multi-camera action dataset for cross-camera action recognition benchmarking, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 187–196.

[28] J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2649–2656.

[29] R. Li, T. Zickler, Discriminative virtual views for cross-view action recognition, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2855–2862.

[30] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, C. Shi, Cross-view action recognition via a continuous virtual path, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2690–2697.

[31] H. Rahmani, A. Mian, 3D action recognition from novel viewpoints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1506–1515.

[32] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1110–1118.

[33] M. Liu, H. Liu, C. Chen, Enhanced skeleton visualization for view invariant human action recognition, Pattern Recogn. 68 (2017) 346–362.

[34] I. Lee, D. Kim, S. Kang, S. Lee, Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1012–1020.

[35] A. Gupta, J. Martinez, J.J. Little, R.J. Woodham, 3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2601–2608.

[36] H. Rahmani, A. Mian, Learning a non-linear knowledge transfer model for cross-view action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2458–2466.
[37] H. Rahmani, A. Mahmood, D.Q. Huynh, A. Mian, HOPC Histogram of oriented principal components of 3D pointclouds for action recognition, in: European Conference on Computer Vision, Springer, 2014, pp. 742–757.
[38] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.
[39] M. Faraki, M. Palhang, C. Sanderson, Log-Euclidean bag of words for human action recognition, IET Comput. Vision 9 (3) (2014) 331–339.
[40] T. Hussain, K. Muhammad, S. Khan, A. Ullah, M.Y. Lee, S.W. Baik, Intelligent baby behavior monitoring using embedded vision in IoT for smart healthcare centers, J. Artificial Intelligence Syst. 1 (2019) 110–124.
[41] L. Liu, S. Wang, B. Hu, Q. Qiong, J. Wen, D.S. Rosenblum, Learning structures of interval-based Bayesian networks in probabilistic generative model for human complex activity recognition, Pattern Recogn. 81 (2018) 545–561.
[42] Y. Liu, L. Nie, L. Liu, D.S. Rosenblum, From action to activity: sensor-based activity recognition, Neurocomputing 181 (2016) 108–115.
[43] Y. Lu, Y. Wei, L. Liu, J. Zhong, L. Sun, Y. Liu, Towards unsupervised physical activity recognition using smartphone accelerometers, Multimedia Tools Appl. 76 (8) (2017) 10701–10719.

Amin Ullah received Ph.D. degree in digital contents from Sejong University, South Korea. He is currently working as a Postdoc Researcher at the Intelligent Media Laboratory, Department of Software, Sejong University, South Korea. His major research focus is on human action and activity recognition, sequence learning, image and video analytics, content-based indexing and retrieval, IoT and smart cities, and deep learning for multimedia understanding. He has published several papers in reputed peer reviewed international journals and conferences including IEEE Transactions on Industrial Electronics, IEEE Transactions on Industrial Informatics, IEEE Transactions on Intelligent Transportation Systems, IEEE Internet of Things Journal, IEEE Access, Elsevier Future Generation Computer Systems, MDPI Sensors, Springer Multimedia Tools and Applications, Springer Mobile Networks and Applications, and IEEE Joint Conference on Neural Networks.

Khan Muhammad received his PhD degree in Digital Contents from Sejong University, Republic of Korea. He is currently working as an Assistant Professor at the Department of Software and Lead Researcher of Intelligent Media Laboratory, Sejong University, Seoul, South Korea. His research interests include intelligent video surveillance (fire/smoke scene analysis, transportation systems, and disaster management), medical image analysis, (brain MRI, diagnostic hysteroscopy, and wireless capsule endoscopy), information security (steganography, encryption, watermarking, and image hashing), video summarization, multimedia, computer vision, IoT, and smart cities. He has filed/published over 7 patents and 100 papers in peer-reviewed journals and conferences in these areas. He is serving as a reviewer for over 70 well-reputed journals and conferences, from IEEE, ACM, Springer, Elsevier, Wily, SAGE, and Hindawi publishers. He acted as a TPC member and session chair at more than 10 conferences in related areas. He is also an Editorial Board Member/Associate Editor for six journals such as the Journal of Artificial Intelligence and Systems and International Journal of Intelligent Networks etc. and Review Editor for the Section "Mathematics of Computation and Data Science" in the journal "Frontiers in Applied Mathematics and Statistics".

Tanveer Hussain acknowledged his degree of Bachelor's in Computer Science from Islamia College Peshawar, Peshawar, Pakistan with Gold Medal distinction in 2017. Currently, he is enrolled in joint Master and Ph.D. program at Sejong University, Seoul, Republic of Korea and serving as a Research Assistant at Intelligent Media Laboratory (IM Lab). His major research domains are features extraction (learned and low-level features), video analytics, image processing, pattern recognition, medical image analysis, multimedia data retrieval, deep learning for multimedia data understanding, single/multi-view video summarization, IoT, IIoT, and resource-constrained programming. He has filed/published several patents and articles in peer-reviewed journals and conferences in reputed venues including IEEE Transactions on Industrial Informatics, Internet of Things Journal, Network Magazine, Elsevier Pattern Recognition, Neurocomputing, Pattern Recognition Letters, Wiley International Journal of Energy Research and International Journal of Distributed Sensors Networks, and Springer Multimedia Tools and Applications. He is a student member of IEEE and providing professional review services in various reputed journals such as IEEE Transactions on Cybernetics and IEEE Transactions on Industrial Informatics. He is serving as an Associate Editor for Journal of Biomedical and Biological Sciences and is editorial board member for Journal of Artificial Intelligence and Systems. For further activities and implementations, visit: https://github.com/tanveer-hussain.

Sung Wook Baik received the B.S degree in computer science from Seoul National University, Seoul, Korea, in 1987, the M.S. degree in computer science from Northern Illinois University, Dekalb, in 1992, and the Ph.D. degree in information technology engineering from George Mason University, Far fax, VA, in 1999. He worked at Datamat Systems Research Inc. as a senior scientist of the Intelligent Systems Group from 1997 to 2002. In 2002, he joined the faculty of the College of Electronics and Information Engineering, Sejong University, Seoul, Korea, where he is currently a Full Professor and the Chief of Sejong Industry-Academy Cooperation Foundation. He is also the head of Intelligent Media Laboratory (IM Lab) at Sejong University. His research interests include computer vision, multimedia, pattern recognition, machine learning, data mining, virtual reality, and computer games.