



Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments



Amin Ullah^a, Khan Muhammad^b, Ijaz Ul Haq^a, Sung Wook Baik^{a,*}

^a Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul, Republic of Korea

^b Department of Software, Sejong University, Seoul, South Korea

HIGHLIGHTS

- Action recognition in online data stream acquired from non-stationary surveillance.
- Efficient CNN model is used for frame-level representation.
- An optimized deep autoencoder is presented for learning sequences and squeezing high dimensional features.
- Investigated a non-linear learning approach for action recognition.
- Iterative fine-tuning of the trained recognition model for newly accumulated data.

ARTICLE INFO

Article history:

Received 1 August 2018

Received in revised form 14 December 2018

Accepted 13 January 2019

Available online 28 January 2019

Keywords:

Big data processing

Action recognition

Online data stream analysis

Optimized deep autoencoder

Convolutional neural network

Machine learning

Non-stationary environment

ABSTRACT

Action recognition is a challenging research area in which several convolutional neural networks (CNN) based action recognition methods are recently presented. However, such methods are inefficient for real-time online data stream processing with satisfied accuracy. Therefore, in this paper we propose an efficient and optimized CNN based system to process data streams in real-time, acquired from visual sensor of non-stationary surveillance environment. Firstly, frame level deep features are extracted using a pre-trained CNN model. Next, an optimized deep autoencoder (DAE) is introduced to learn temporal changes of the actions in the surveillance stream. Furthermore, a non-linear learning approach, quadratic SVM is trained for the classification of human actions. Finally, an iterative fine-tuning process is added in the testing phase that can update the parameters of trained model using the newly accumulated data of non-stationary environment. Experiments are conducted on benchmark datasets and results reveal the better performance of our system in terms of accuracy and running time compared to state-of-the-art methods. We believe that our proposed system is a suitable candidate for action recognition in surveillance data stream of non-stationary environments.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Human action recognition encompasses many important domains of real-life such as intelligent videos surveillance, detection of abnormal and suspicious actions, video retrieval based on different actions, video semantic recognition, and patients monitoring in healthcare centers [1–3]. There are numerous applications of action recognition using online data stream such as monitoring through visual sensors in surveillance, videos from websites, and social media feeds, that can lead to detect initiated anomaly, fraud or any abnormal situations [4]. In the context of videos, human actions can be recognized by the movement of different body parts such as hands and legs. A single still image cannot convey the

whole idea of an action [5]. For example, jumping for a head-shot in football and skipping rope have the same action pose in the initial frame. The discrimination of both actions can be captured in a sequence of frames. Analyzing the movements of a human body in frame sequence and interaction with surrounding leads to recognizing the perfect actions in the video data stream [6,7].

In non-stationary data streams whenever variation in new data is encountered, the trained model over the previous data cannot be considered effective enough. The reason is its adaptability issue over the new distribution of data which needs diversity for non-stationary environment [8]. To overcome this issue, Lobo et al. [9] considered it as an optimization problem, which is solved by a bio-inspired algorithm to validate the heterogeneity of drifts and achieved high diversity through self-learning optimization technique. Another novel approach is proposed by Krawczyk et al. [10] modified weighted one-class SVM and improved it for the non-stationary streaming data analysis. They claimed that one-class

* Corresponding author.

E-mail addresses: qamin3797@gmail.com (A. Ullah), sbaik@sejong.ac.kr (S.W. Baik).

classifier can adapt its decision boundary according to new data streams along with forgetting mechanism which helps the model to re-learn the parameters. Similarly, Bartosz et al. [11] presented an efficient ensemble learning technique for recognizing activities in real-time. The system iteratively modifies the weights of Naïve Bayes classifier and make it smoothly adaptable to current situation of stream even without an external drift detector. Abdallah et al. [12] presented a detailed survey about activity recognition in online data stream mining. Moreover, recognizing human actions accurately in real-time from online surveillance data stream is a highly challenging task due to computation of high-dimensional features, variation in viewpoint, motion, cluttered backgrounds, occlusion, and different illumination conditions [13–15].

To address these problems, numerous handcrafted local feature descriptors were used in the domain of recognizing human actions since the last decade [16–18] in which the number of spatial–temporal based approaches were significant [18–20]. Such schemes are based on the analysis of motion information and can be improved in performance by Bag-of-Words (BoW) [21–23] but developing BoW is computationally expensive and requires hard engineering. Dalal et al. [24] presented motion boundary histogram (MBH), where edges motion are captured in HOG descriptor. Local gradients in horizontal and vertical components of the optical flow are calculated separately. The corresponding magnitude and orientation in both components are used as weighted votes for local orientation histograms. The extended version of HOG feature descriptor named HOG3D is presented by Klaser et al. [17] in which 3D gradients orientation computed from integral video representation are binned into polyhedrons to analyze appearance and motion information. This scheme has expensive quantization cost due to high dimensionality structure. To handle this problem, Shi et al. [25] introduced Gradient Boundary Histograms (GBH). They used time-derivatives of image gradients instead of simple gradients to highlight the moving edge boundaries. Klaser et al. [26] proposed Optical Flow Co-occurrence Matrix (OFCM) to extract a set of statistical measures captured using the magnitude and orientation of optical flow. The key motivation to design OFCM was on assumption that the spatial relationship can be found in the local neighborhoods of the flow field, which has a major contribution in the representation of motions.

Handcrafted features extraction mechanisms involve hard engineering, represents low-level semantic of visual data, and high complexity for extraction and classification. Therefore, automatic features learning methods are initiated by researchers. For instance, neural networks-based methods can directly extract features from raw inputs based on its trained weights and biases. CNN learns features in a hierarchical way where initial layers acquire local features from visual data and the final layers extract global features representing high-level semantics [27]. Recently, researchers have tried to develop a variety of CNN architectures for sequence learning of action recognition. For instance, a CNN framework based on spatio-temporal information is proposed by Karpathy et al. [28] to learn the motion features. Several temporal information fusion schemes are analyzed to fuse local motion direction with global features. However, the recognition rate is 63.3% on UCF 101 dataset [29], indicating that their CNN architecture is unable to effectively represents human actions in the video stream. In another work, Park et al. [30] extracted motion information for a specific part of the image using a spatial network, that captures highly activated features from magnitude information of the optical flow. Features maps of the last convolutional layer of spatial network are analyzed to compute optical flow magnitudes. Another similar work presented by Simonyan and Zisserman [31] is based on a two-stream network. The first stream involves spatial network which extracts temporal information from the sequence of frames. In the second stream, temporal network is utilized to

compute dense optical flow displacements across multiple frames. Finally, the average scores from both streams are used for prediction. Majority of the deep CNN frameworks are for 2D images without keeping time information. Ji et al. [32] presented a 3D CNN for end to end action recognition. Their model extracts features from both the spatial and temporal dimensions to get motion information from multiple adjacent frames. This method is based on analyzing consecutive segments of human subjects in video frames.

Deep CNN based methods can learn influential weights to discriminate between different actions present in visual data [33]. However, action recognition models are not trained on a large-scale dataset such as ImageNet. Many studies [6,27,34] have concluded that the activations of pre-trained CNN models achieved impressive success for image retrieval, fire detection, and video summarization. Therefore, we have extensively investigated the deep features of various pre-trained CNN models for action recognition. Furthermore, existing CNN models are computationally expensive and their recognition accuracy is not satisfied for all environments such as online data streams of non-stationary environment. Therefore, we conduct this study to address these issues with the following key contributions:

1. We propose an efficient and optimized action recognition system to process data streams acquired from visual surveillance of non-stationary environment. Our system uses the activations of fully connected layer of a pre-trained VGG-16 CNN model for frame level representation of an action in video streams.
2. Actions are sequence of motion patterns in consecutive frames, and the frame level features are high-dimensional raw data to recognize actions precisely. Therefore, we have trained an optimized DAE to squeeze those features and make it able to associate frame to frame hidden changes in low-dimensional feature plane. This enables our system in effective sequence learning for action recognition compared to complex learning approaches such as long short-term memory (LSTM).
3. We have investigated a non-linear learning approach and trained an efficient quadratic SVM to recognize actions from low-dimensional features plane.
4. The video data in non-stationary environments are very diverse in nature due to changing overtime, where one-time trained models are not effective enough for precise predictions. Therefore, we introduce an iterative fine-tuning module that collects new data of high confidence prediction for actions and iteratively fine-tune the recognition model with this data. This process makes our system capable of updating itself according to the variations in the underlying non-stationary environment.
5. Our system is tested on benchmark datasets from different perspectives and results are encouraging compared to state-of-the-art, making it suitable for real-time surveillance monitoring in general and data streams of non-stationary environment in particular.

The remaining paper is organized as follows: Section 2 highlights various aspects of the proposed framework. Experimental setup and discussion on results are given in Section 3. Conclusion, strengths, weaknesses, and future directions of this work are discussed in Section 4.

2. Proposed framework

In this section, the mechanism of the proposed system is discussed in detail. The system includes representation of actions in online video data stream using deep CNN features, sequence

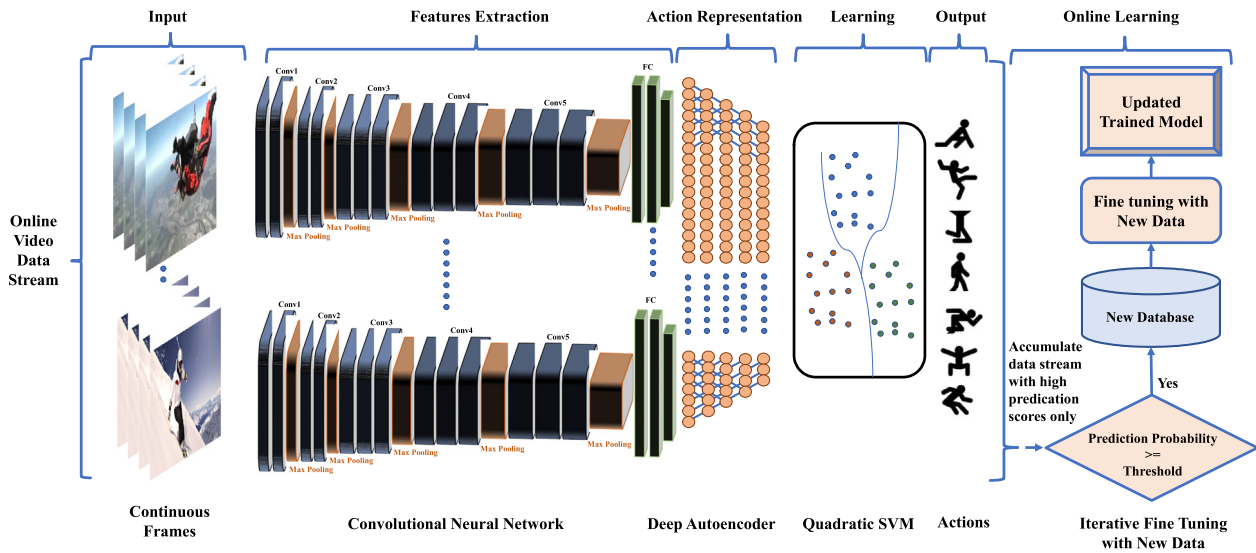


Fig. 1. The proposed framework for human action recognition system. An online video data stream is forward propagated to a pre-trained CNN model to extract features from a fully connected layer. This is followed by a deep autoencoder which learns the temporal changes of a human action in low dimensional features plane. Finally, a quadratic SVM is used to classify human actions.

Algorithm 1: Human actions representation and recognition in online data stream

Input: Video stream V

Initialization

- 1) Load trained models (m_1 : VGG-16 model, m_2 : deep autoencoder, and m_3 : quadratic SVM)
- 2) Select duration \mathcal{D} (in secs) of online video stream for processing
- 3) Set counter=1

Process

- 4) **while** (V)
 - a. $\mathcal{D} \leftarrow \text{read } V[\text{counter}]$
 - b. $f_1 \leftarrow m_1(\mathcal{D})$
 - c. $f_2 \leftarrow m_2(f_1)$
 - d. $P_{\text{counter}} \leftarrow m_3(f_2)$
 - e. Predicted action in video data stream for duration $\mathcal{D} \leftarrow P_{\text{counter}}$
 - f. **If** $P_{\text{counter}} > \text{threshold}$
 - Accumulate data of \mathcal{D} with P_{counter} in new database
 - end of **If**
 - g. Increment counter
 - h. **Output:** display label of an action for \mathcal{D}
- 5) end of **while**

scores of different actions and iteratively fine-tune our recognition model with new data to adopt variations of non-stationary environment. The proposed framework is shown in Fig. 1 whereas each phase of the system is described in a different section and the implementation steps are given in Algorithm 1.

2.1. Deep features extraction and preparation

Video data contain a large amount of hidden visual contents including temporal changes of texture, motion, edges, and colors. An efficient representation and analysis of these contents allow us to make automatic timely decisions such as human actions recognition, fire detection, contextual information analysis, and event detection. Deep neural networks have shown its effectiveness in images [35], sounds, and videos analysis [36] because of its remarkable representational abilities [37]. Training a deep CNN model requires a huge amount of data and high-cost computational resources. The solution to this problem is to use pre-trained CNN models for different problems as their parameters are trained on enormous datasets such as ImageNet [38]. In the proposed system, a fully connected “FC8” layer of a pre-trained VGG-16 [39] CNN model is used to extract features from video frames. The fully connected layer extracts generic global descriptors from an image [34]. Therefore, we argue that these features are highly dominant and capable of representing visual contents which can help us for learning the complex sequences in the video frames. The pre-trained CNN model process one frame at a time, where the video data contains a sequence of frames. We have fed 15 frames to the employed CNN model taken from online video data stream of one-second with one frame skip. It gives a high dimensional features vector representing human action in a raw form. The temporal sequences between these features are squeezed through an efficient DAE and learned it using a quadratic SVM for human action recognition.

2.2. VGG-16 CNN model

In the proposed framework a VGG-16 [39] CNN architecture is chosen for deep features extraction from video frames. Because it is noticed in our experiments, that it can achieve sensible stability

learning using DAE, and classification of actions with a quadratic SVM. First, deep features are extracted from selected frames of online data stream using a pre-trained VGG-16 CNN model. Second, the high dimensional features are squeezed and the temporal changes between features are learned in a low dimensional feature plane using DAE. Finally, a quadratic SVM is trained to classify the squeezed features of processed duration of online video stream. Furthermore, we accumulate the data stream with high confidence

Table 1

The architecture of a pre-trained VGG-16 CNN model.

Layers	Conv1a Conv1b	Conv2a Conv2b	Conv3a Conv3b	Conv3c	Conv4a Conv4b	Conv4c	Conv5a Conv5b	Conv5c	FC	FC	FC8
Kernel size	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	1 × 1	3 × 3	1 × 1	Inner product	Inner product	Inner product
Stride, Pad	1, 1	1, 1	1, 1	1, 1	1, 1	1, 1	1, 1	1, 1	4096	4096	1000
Channels	64	128	256	256	512	512	512	512			
	Max pooling	Max pooling		Max pooling		Max pooling		Max pooling			

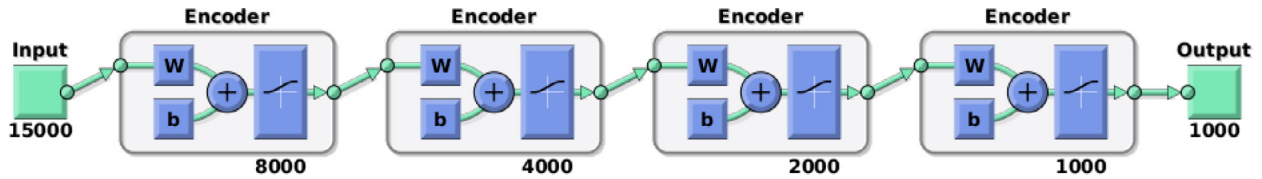


Fig. 2. The architecture of the stacked four autoencoders squeezing high-dimensional features to low-dimensional features.

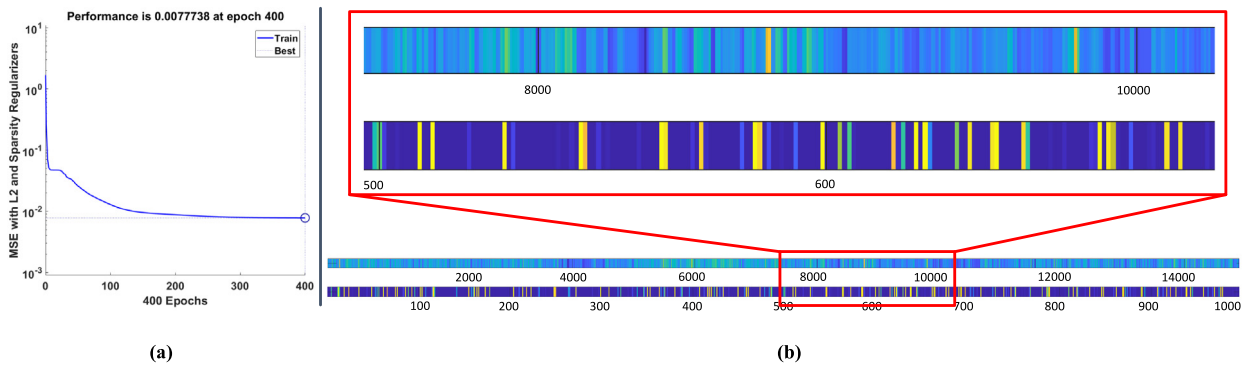


Fig. 3. Performance of the feature transformation from high-dimension to low-dimensions. (a) Mean square error graph of DAE decreasing with training epoch. (b) Comparison between the encoded features with some part of the original features.

between the accuracy and time efficiency for action recognition problem. The architecture of VGG-16 is given in Table 1. The architecture of VGG-16 is different from previous state-of-the-art CNN models [40–43] where initial layers are convolved with 11×11 or 7×7 kernels with 4 to 5 strides. This type of setup increases the number of parameters in a CNN model. Furthermore, with wide stride, it can miss important patterns in the image. On the other hand, VGG-16 contains 3×3 kernels for all convolutional layers with 1 stride, helping to reduce the number of parameters in layers and convolve each pixel of the image due to 1 stride. It can be seen from Table 1 that two consecutive 3×3 convolution layers are applied without a pooling layer in between. This combination of two-layer results in effect of 7×7 kernels. Assembling consecutive three convolutional layers are followed by a “relu” activation, where multiple non-linear functions make it more discriminative.

2.3. Deep autoencoder

Deep autoencoder is an effective unsupervised feature representation technique with multiple hidden layers. The motivation behind the neural concept of data learning is that the parameters of hidden layers are not manually constructed [44,45], but they are learned according to the given data automatically. This idea encouraged us to learn time axis features of video sequences using DAE. The high dimensional deep features are squeezed to low dimensions with a negligible error during transformation. Deep features from the sequence of frames are extracted and learned its hidden patterns and frame to frame changes using an efficient

architecture of four layers stacked autoencoder as shown in Fig. 2. The first layer encodes 15000-dimensional feature vector to 8000 neurons, pursued by 4000, 2000, and 1000 dimensions reduction, respectively. The reason behind reduction of high dimensional data with half factor is to reduce time complexity of the autoencoder. Squeezing high dimensional data with small steps and multiple deep layers results in high computational complexity. The DAE learns “hierarchical grouping” or “part-whole decomposition” in the input data [46]. The initial layers of the stacked autoencoder capture the first order features and changes in the raw input data. On the other hand, the intermediate layers learn the second-order feature corresponding to the patterns that come in the first-order features. Therefore, we argue that the proposed DAE learns the changes and patterns of human action in video sequence effectively.

The autoencoder is comprised of two phases: First is encoding where data is multiplied by weights, biases are added, and followed by some non-linearity function such as sigmoid and relu given in Eq. (1). Secondly, the encoded data is decoded to the same number of inputs as shown in Eq. (2). The weights are adjusted using a backpropagation to reduce the mean squared error near to zero.

$$h(x) = \text{sigm}(Wx + b) \tag{1}$$

$$\hat{x} = \text{sigm}(W(h(x)) + b) . \tag{2}$$

In the stacked autoencoder, the first hidden layer takes input x , while the other gets input from the previous hidden layer in the network as shown in Eqs. (3) and (4). Herein, “ n ” is the number of

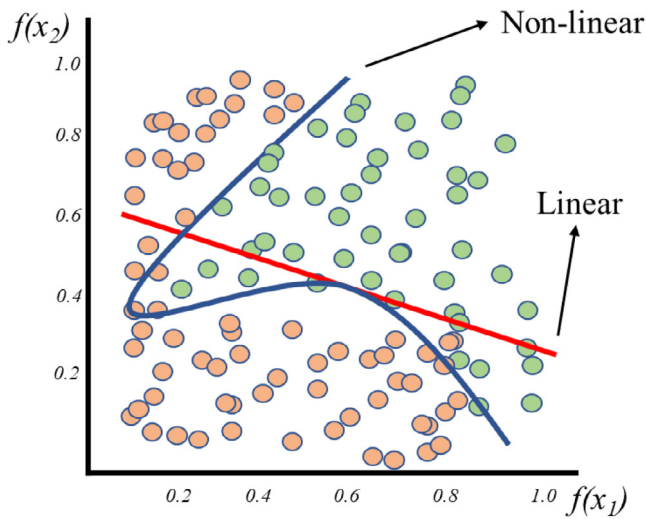


Fig. 4. Hyperplane separation between two class data using linear SVM and non-linear SVM classification.

layers for encoding, x^l , W^l , and b^l are the data, weights, and biases of the concerned layer, respectively.

$$h(x)^{(l+1)} = \text{sigm}(W^l x^l + b^l) \tag{3}$$

$$\hat{x}^{(n+1)} = \text{sigm}(W^{(n-1)} h(x)^{(n+1)} + b^{(n-1)}) \tag{4}$$

The proposed DAE is trained up to 400 epochs. The L2 weights regularization is applied to diminish the over-fitting problem and “falling into local minima” problem. The sparsity regularization is also applied with sigma α of value 0.05 which means that each neuron in the hidden layer takes an average output of 0.5 over the training samples. Finally, mean squared error (MSE) with L2 regularization and sparsity regularization is used as a cost function for fine-tuning the weights of the DAE. The error is reduced up to 10^{-2} in 300 epochs, which was 0.0077 for the last epoch of the

training phase. Fig. 3(a) shows performance graph of the training phase, where we can see that the error is reduced smoothly without going into overfitting problem. Fig. 3(b) represents the comparison between the encoded data with some portion of the original data. It can be noted that data having low activation, are caught by the sparsity regularization and high values have got the same graph pattern as the original data.

2.4. Learning actions using quadratic SVM

Learning with linear SVM is not effective in a high dimensional features plane when substantial class overlapping comes in the training data as the hyperplane separating two different classes, is always a straight line [47] as shown in Fig. 4. In such case, a non-linear SVM is effective which can separate the data with wide hyperplane between two class data. In training SVM for multiple classes, we get the imbalance data problem and we need to train one category data against all categories, because SVM is originally for binary class classification. In the non-linear SVM, increasing the polynomial function may give an optimum hyperplane between the two classes, but it increases the computation time of the system [48]. In our system, we have used a non-linear quadratic SVM through which we have achieved stability between accuracy and time efficiency.

$$M = \frac{N(N-1)}{2} \tag{5}$$

Two strategies have been used for multi-class SVM: one-vs-one (OvO) and one-vs-all (OvA). In “OvO”, we need to train “M” classifiers for “N” classes as given in Eq. (5) where the n th class is trained against all $N - 1$ classes. This is computationally expensive when the number of classes increases. In case of the proposed system, it is trained on UCF101 [29], HMDB51, and UCF50 datasets, which has 101, 51, and 50 categories, respectively. This lead to 5050 classifiers for 101 categories of data, 1275 and 1225 classifiers for 51 and 50 classes, respectively. This type of setup is not efficient in real-time applications such as surveillance stream. On the other hand, in “OvA” we need only “N - 1” classifiers. It requires training

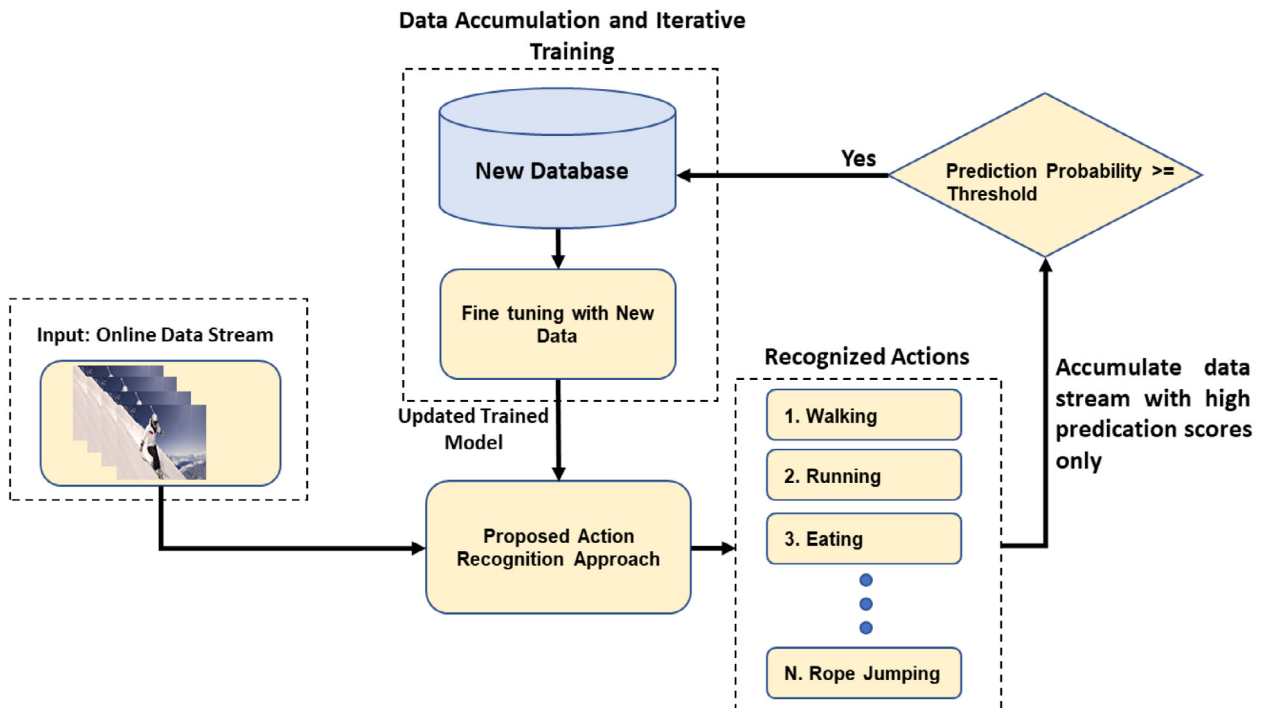


Fig. 5. Iterative fine-tuning process of the proposed system.



Fig. 6. Sample action classes form UCF101, UCF50, HMDB51, and YouTube actions datasets.

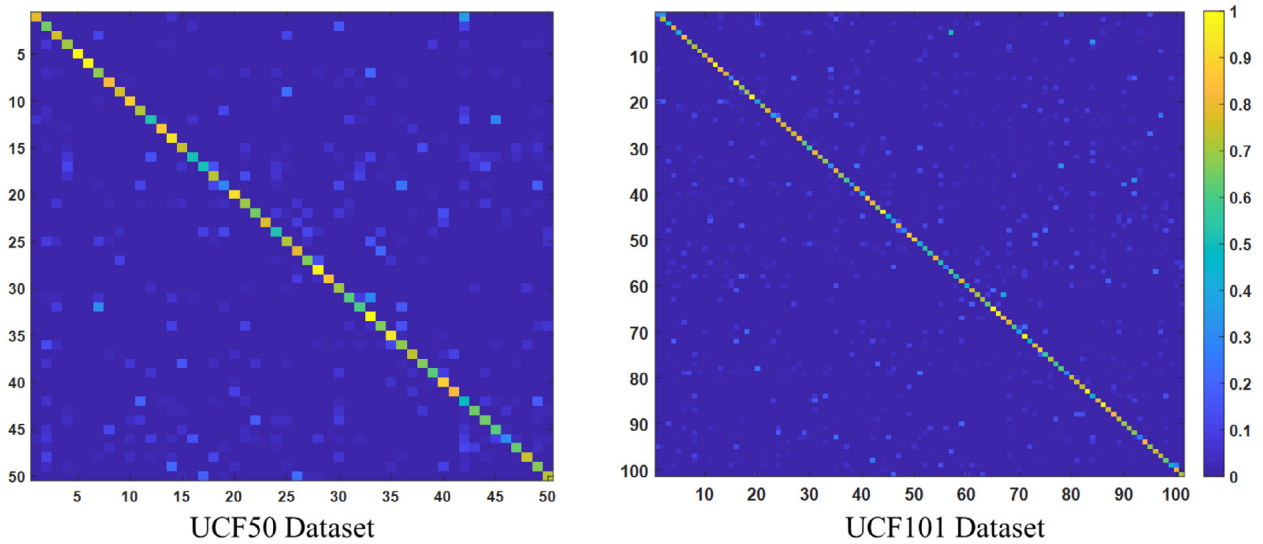


Fig. 7. Confusion matrixes of UCF50 and UCF101 action recognition datasets.

of single classifier per-class, with the training data of one class as positive samples and all other classes data as negatives samples. In such case, the classification gets into imbalance data problem due to samples of one class data against samples of all classes. In the proposed system, we have used a highly powerful feature which is discriminative for each class, therefore, we get high accuracy on “OvA” non-linear classifier.

2.5. Incremental model training for data stream

Another encouraging feature of our system is its capability to update itself iteratively, grasping the changes in surrounding environment. The iterative training process is shown in Fig. 5. The data stream predicted with confidence score greater than a certain threshold can be used to re-train the model iteratively, enabling it adaptive to different varying conditions. The threshold can be selected considering the requirements of users, environmental effects, and deployment settings. The data with high confidence scores is accumulated along with the predicted labels. When a certain amount of data is assembled, the same trained model is

fine-tuned with new data which adapts itself with the variations in the environment. Considering these characteristics, our system can be implemented in health care centers for patient activity monitoring and for anomaly, fraud and intruder detection in real-time video streams of surveillance.

3. Experimental evaluation

In this section, we evaluated the proposed system on several benchmark datasets used for action recognition assessment. The datasets include UCF101 [29], UCF50 [58], YouTube Actions [59], and HMDB51 [60]. Sample categories from five different datasets are shown in Fig. 6. The proposed system was implemented and assessed using MATLAB2017b on Ubuntu16.04 environment, Core™i5-6600 set up with 16 GB RAM and 12 GB GeForce-Titan-X GPU. A deep learning toolbox “MathConvNet” is used for CNN features extraction, neural network toolbox is used for DAE, and “classification Learner” toolbox is used for learning quadratic SVM classifier for action recognition. The proposed method is assessed

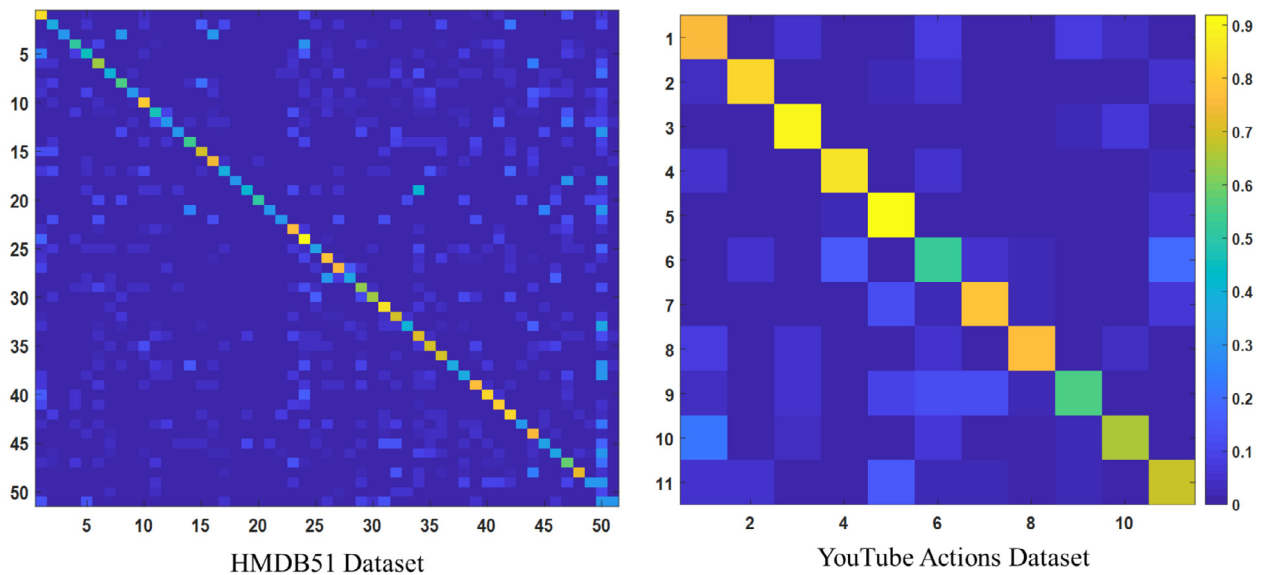


Fig. 8. Confusion matrixes of HMDB51 and YouTube action recognition datasets.

Table 2

Comparison of the proposed system with state-of-the-art hand-crafted and CNN based action recognition methods.

Methods		(Accuracy %)			
		UCF50	UCF101	HMDB51	YouTube action
Hand crafted features-based methods	HOG–HOF + FV [49,50]	–	75.4	45.6	–
	OHF + BoW [24,50]	–	77.1	51.5	–
	GBH + MBH [25]	–	86.6	62.2	–
	Improved dense trajectories hybrid approach [21]	92.3	87.9	61.1	–
	Multi-view super vector [51]	–	83.5	55.9	–
Neural network-based methods	LSVC with CNN [28]	–	65.4	–	–
	Composite LSTM [52]	–	75.8	44.1	–
	Hierarchical clustering [53]	93.2	76.3	51.4	89.7
	key-volume mining deep CNN [54]	–	93.1	63.3	–
	F_{STCN} (SCI fusion) [55]	–	88.1	59.1	–
	Fusion (S: VGG-16, T:VGG-16) [56]	–	92.5	65.4	–
	Single stream CNN [57]	–	–	–	93.1
	Two-stream model (fusion by SVM) [31]	–	88.0	59.4	–
	Proposed system	96.4	94.33	70.3	96.21

using four different accuracy calculation metrics including confusion matrix (Fig. 7 and Fig. 8), precision, recall, and the F-measure score given in Table 3. Results of each dataset and comparisons with state-of-the-art are given in Table 2 and discussed in separate sections.

3.1. UCF101 dataset

UCF101 [29] is a popular real-life action videos dataset which contains 13320 videos of 101 action classes collected from YouTube in “.avi” format. The number of samples for all categories are balanced, ranging from 100~130 samples and duration of an action in each sample is 2~7 s. UCF101 is relatively a challenging dataset because of many action classes where humans are performing actions on different objects, musical instruments, sporting goods, and interaction with parts of human body. Results are collected using this dataset and comparison with state-of-the methods is given in Table 2, where column 4 presents the percentage accuracy of the proposed system. Confusion matrix for the test set on this dataset is shown in Fig. 7 where we have achieved an overall accuracy of 94.33%. We compared the proposed system with the

histograms of oriented gradient and optical flow (HOG–HOF) + fisher vector (FV) [13], oriented histogram flow (OHF) + bag of words (BoW [15], gradient boundary histograms (GBH) + motion boundary histogram (MBH) [25], improved dense trajectories (IDT) hybrid approach [21], and multi-view super vector [51] hand-crafted features. The IDT [51] shows the best accuracy of 87.9% on UCF101 dataset among all hand-crafted features based methods. We have outperformed these methods by 6.43% increase in accuracy, which is evident from column 4 of Table 2. The performance of the proposed system beat neural network-based methods [28, 31,52–56] with 1.23% improvement in the accuracy. The precision, recall, and F1-measure score for the UCF101 dataset is given in Table 3, where we have achieved positive prediction score 0. 8932, 0. 9035 sensitivity score, and 0. 8983 F1-measure scores. These statistics show the effectiveness of the proposed system for human action recognition on the UCF101 dataset.

3.2. UCF50 dataset

UCF50 is a diverse collection of human actions due to high diversity in camera motion, object appearance and pose, object

scale, viewpoint, cluttered background, and different illumination in surroundings [58]. It has fifty action classes, wherein videos of each category are divided into different groups that share some common features variations such as in one group a piano is played by a person four times but with a different viewpoint. The comparison with state-of-the-art methods on this dataset is given in Table 2. It is evident from column 3 of Table 2 that the proposed system has achieved higher accuracy compared to both hand-crafted features and deep features-based methods. Confusion matrix of the test set for UCF50 dataset is shown in Fig. 7 where we have achieved percentage accuracy of 96.4%. The proposed system has reported an improvement of 4% in results to hand-crafted features-based method (IDT) [21] and 3% to CNN features based hierarchical clustering [53] method. The precision, recall, and F1-measure scores for the UCF50 dataset are given in Table 3. The proposed method has achieved higher true positive, sensitivity and effectiveness scores 0.9321, 0.9124, and 0.9221, respectively.

3.3. HMDB51 dataset

The HMDB51 [60] is considered as a challenging dataset in action recognition society. It contains actions related to human facial interaction, motion of the body parts, body dealing with objects, sports, and human exercises. The dataset comprises of 6849 action samples taken from YouTube with a variety of different subjects and is divided into 51 categories. In each category there are more than one hundred video samples. The dataset is made more challenging because of taking samples from a variety of subjects in different illumination and viewpoint changes for performing the same actions. On this dataset, the accuracy of state-of-the-art methods lies under 60%. In the proposed system, we represented human action with high-level features using CNN and

Table 3

The effectiveness of the proposed system using precision, recall, and F1-measure scores.

Dataset	Precision	Recall	F1-measure
UCF50 [58]	0.9321	0.9124	0.9221
UCF101 [29]	0.8932	0.9035	0.8983
HMDB51 [60]	0.6906	0.6234	0.6553
YouTube actions [59]	0.9541	0.9387	0.9463

DAE, which helps to recognize the complex action in an efficient way with improved accuracy. The comparison with previous techniques reported for HMDB51 is given in Table 2, where the proposed system has dominated [21,25,51]. The method [25] shows best performance of 62.2% accuracy on HMDB51 among all hand-crafted features based methods. The proposed system is also compared with CNN based methods including composite LSTM [52], hierarchical clustering [53], key-volume mining deep CNN [54], FSTCN (SCI fusion) [55], fusion (S: VGG-16, T:VGG-16) [56], and two-stream model (fusion by SVM) [31]. The proposed system has improved performance of these methods by 8.2% in accuracy of hand-crafted features and 5% for the CNN based approaches. Confusion matrix for HMDB51 dataset is shown in Fig. 8 where we have achieved percentage accuracy of 70.4%. The precision, recall, and F1-measure score for HMDB51 dataset is given in Table 3. On this dataset, we have achieved 0.6906 precision, 0.6234, and 0.6553 for recall and F1-measure score, respectively. The low recall score is due to high false positives received for some actions such as Fig. 8 (class 50), which has comparatively many false positives. Despite these factors, our system achieved a higher accuracy of 70.4% on this challenging dataset.





Sequence of frames representing an action	Ground Truth	Predictions	Confidence Score
	Basketball	Basketball	0.61
	Skate Boarding	<u>Skiing</u>	<u>0.231</u>
	Surfing	Surfing	0.41
	Horse Racing	Horse Racing	0.74
	Diving	<u>High Jump</u>	<u>0.24</u>
	Skiing	Skiing	0.76

Fig. 9. Predicted results along with maximum confidence scores for the overall test video by the proposed action recognition system. The underline red text shows in-correct prediction of our system.













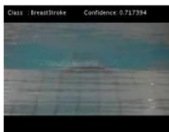



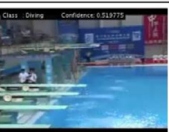



Intervals	Int. 1	Int. 2	Int. 3	Int. 4	Int. 5	Overall Accuracy
Representative frame from each interval						100%
GT/Prediction	Horse Ride / Horse Ride	Horse Ride / Horse Ride	Horse Ride / Horse Ride	Horse Ride / Horse Ride	Horse Ride / Horse Ride	
Confidence score	0.527	0.452	0.485	0.469	0.526	
Representative frame from each interval						100%
GT/Prediction	Baseball / Baseball	Baseball / Baseball	Baseball / Baseball	Baseball / Baseball	Baseball / Baseball	
Confidence score	0.500	0.483	0.333	0.333	0.385	
Representative frame from each interval						80%
GT/Prediction	Breast Stroke / Biking	Breaststroke / Breaststroke	Breaststroke / Breaststroke	Breaststroke / Breaststroke	Breaststroke / Breaststroke	
Confidence score	0.348	0.682	0.717	0.644	0.717	
Representative frame from each interval						80%
GT/Prediction	Dive / Jump	Dive / Dive	Dive / Dive	Dive / Dive	Dive / Dive	
Confidence score	0.312	0.519	0.549	0.509	0.492	

Fig. 10. Sample variations in predictions and confidence scores for an action with respect to time. The overall accuracy is considered from the predictions of the particular action for all five intervals.

3.4. YouTube actions dataset

YouTube action is one of the complex and challenging datasets because the actions samples are collected in low resolution with moving and still camera in different scales, clustered background, illumination and viewpoint changes. It contains 11 action classes collected from sports in which some videos are taken from 25 subjects with 4 samples for each action and other videos are collected from YouTube. The comparative results using this dataset with other methods is given in Table 2. Two CNN based methods including hierarchical clustering [53] and single stream CNN [57] achieved 94.3% and 93.1% accuracy, respectively, which is improved by our proposed method by almost 2%. Confusion matrix of the test set for YouTube actions dataset is presented in Fig. 8 where we have achieved percentage accuracy of 96.21%. The proposed system exhibits better precision, recall, and F1-measure score of 0.9541, 0.9387, and 0.9463, respectively for this dataset as given in Table 3.

3.5. Discussion on visual results

The proposed system is evaluated on 20% samples from each dataset to obtain the quantitative results. Fig. 9 shows correct and incorrect predictions with their maximum confidence scores of our proposed system for a single action video. A set of frames of each action are also given for better understanding of readers. It is noted from the experiments that the complex actions such as

“Horse racing” and “Horse riding” which has minor difference of “many horses” and “one horse”, respectively, has more than 95% accuracy. Row 2 and Row 5 of Fig. 9 show incorrect predictions. This is due to the common visual contents in the ground truth and predicted classes of the dataset. For example, “skateboarding” is predicted as “skiing” and “diving” is miss-classified as “high jump”. However, the confidence scores of incorrect predictions of these challenging classes are very low. The predication “high jump” is incorrect because when the diver jumps for the dive, changes between the frames, represent high jump class for that particular interval of time.

3.6. Model behavior vs. label transitions over time

The proposed trained model predicts action in the data stream over intervals, and for each interval of time, the confidence score changes due to motion of the camera, variation in the viewpoint, and scale of an actor. It is very challenging to have adjacent prediction scores for the same action or abrupt change during transition to another action. A series of confidence scores and predicted actions over time are visualized in Fig. 10, where scores for the same class are approximately similar to each other. However, when there are some overlapping frames of two actions in the interval under process, it affects the accuracy of prediction in a negative manner with a low confidence score. This situation can be observed from Fig. 10 (Row 3 and Row 4). For instance, in Row 4 the frames of “jumping” action are overlapped with frames of “dive”

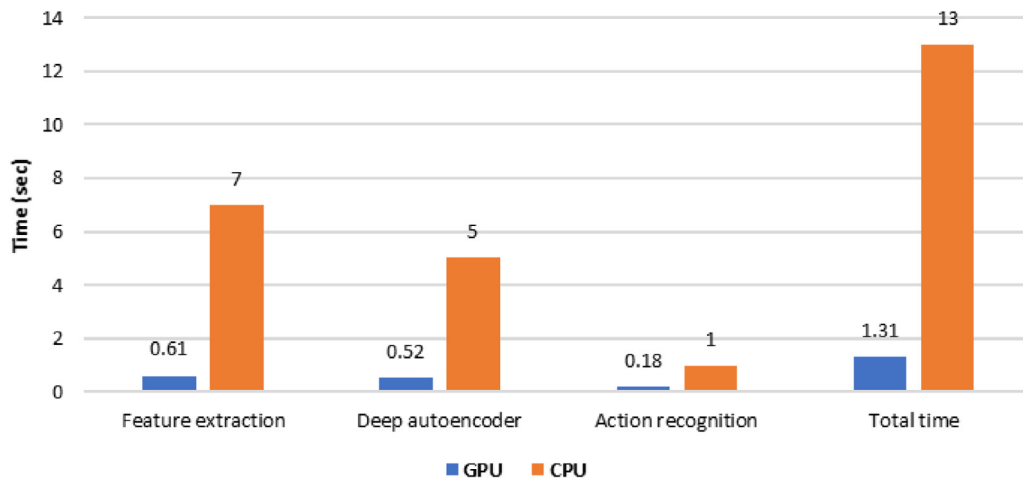


Fig. 11. Processing time of the proposed system on CPU and GPU taken for one second video data stream for action recognition.

action. Therefore, in the first interval of the concerned sequence, high jump is predicted with low confidence score and for rest of the intervals, scores are adjacent and predictions are accurate. This problem is tackled by iteratively fine-tuning the recognition model, which makes it adoptable to variations between different actions performed in non-stationary environment.

3.7. Computational complexity and feasibility analysis

This section investigates the running time of our system and its feasibility to online data stream understanding and mining. The experiments of the proposed system for feature extraction, training, and testing of DAE and quadratic SVM for action recognition are performed using GeForce-Titan-X GPU. On this setup, the VGG-16 model takes 0.12 s to extract features from one frame. In the proposed system, we have fed 15 frames at a time from video data stream to take advantage of the parallelism in GPU which takes approximately 0.61 s for extracting features from them. Secondly, the DAE takes 0.52 s for squeezing action patterns to low dimensional features plane. Finally, the quadratic SVM takes only 0.18 s to classify the given set of frames into action categories. As a whole, the system takes 1.31 s for processing 30 frames from the video stream, showing nearly real-time processing. Based on the statistics shown in Fig. 11, the proposed system can process 25 frames per second in real-time surveillance for human action recognition in non-stationary environments.

4. Conclusion and future work

In this paper, we presented an optimized DAE based human actions representation framework that can be implemented in real-world dynamic scenarios. The input of our system can be acquired from online surveillance video data stream, websites, social media feeds or any other visual content resources. Semantic features of a pre-trained VGG-16 CNN model are used for frame level representation. An optimized DAE is trained to effectively represent actions from raw information of video frames. The DAE converts high-dimensional data to low-dimensional feature plane and learns information variations amongst adjacent frames. Finally, quadratic SVM processes the output of DAE to classify the human action performed within the input video data stream at a particular time. Our experiments verify that the proposed system can process 25 frames per-second regardless of the noisy effects and heterogeneous nature of data streams. The experiments conducted with benchmark datasets including UCF50, UCF101, HMDB51, and

YouTube Action dataset revealed that it is an efficient and effective system for action recognition in surveillance from non-stationary environment. Lastly, data stream with high confidence scores are accumulated for iterative fine-tuning of the proposed action recognition model with new data, enabling it to adopt variations in non-stationary environment.

In future, we plan to analyze multiple actions by detecting and tracking multiple targets in a sequence of online video stream. The current available realistic video datasets contain actions performed by a single person, where multiple actions need to be recognized in dense situations. Secondly, when there is a situation of overlapping actions in a single sequence such as jump and dive in same sequence reduces the accuracy. This limitation will be considered in our future work. Furthermore, we have motivation to develop action recognition mechanism based on multi view surveillance videos connected in a visual sensor network in different dynamic environments. Finally, the proposed system is feasible to be extended for video classification, human activity recognition, violent event recognition, and can be implemented for crowd analysis in dense environment.

Acknowledgment

This work was supported by the National Research Foundation of Korea Grant funded by the Korea Government (MSIP) under Grant 2016R1A2B4011712.

References

- [1] L. Liu, et al., Recognizing complex activities by a probabilistic interval-based model, in: AAAI, 2016.
- [2] M. Luo, et al., An adaptive semisupervised feature analysis for video semantic recognition, IEEE Trans. Cybern. 48 (2) (2018) 648–660.
- [3] Y. Liu, et al., From action to activity: sensor-based activity recognition, Neurocomputing 181 (2016) 108–115.
- [4] S. Luo, et al., Action recognition in surveillance video using convnets and motion history image, in: International Conference on Artificial Neural Networks, Springer, 2016.
- [5] Y. Liu, et al., Fusion of low-and high-dimensional approaches by trackers sampling for generic human motion tracking, in: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE, 2012.
- [6] A. Ullah, et al., Action recognition in video sequences using deep Bi-directional LSTM with CNN features, IEEE Access (2017).
- [7] A. Ullah, et al., Activity recognition using temporal optical flow convolutional features and multi-layer LSTM, IEEE Trans. Ind. Electron. (2018).
- [8] J.L. Lobo, et al., DRED: An evolutionary diversity generation method for concept drift adaptation in online learning environments, Appl. Soft Comput. 68 (2018) 693–709.

- [9] J.L. Lobo, et al., On the creation of diverse ensembles for nonstationary environments using Bio-inspired heuristics, in: International Conference on Harmony Search Algorithm, Springer, 2017.
- [10] B. Krawczyk, M. Woźniak, One-class classifiers with incremental learning and forgetting for data streams with concept drift, *Soft Comput.* 19 (12) (2015) 3387–3400.
- [11] B. Krawczyk, Active and adaptive ensemble learning for online activity recognition from data streams, *Knowl.-Based Syst.* 138 (2017) 69–78.
- [12] Z.S. Abdallah, et al., Activity recognition with evolving data streams: A review, *ACM Comput. Surv.* 51 (4) (2018) 71.
- [13] Y. Wang, G. Mori, Hidden part models for human action recognition: Probabilistic versus max margin, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (7) (2011) 1310–1323.
- [14] Y. Liu, et al., Action2Activity: Recognizing complex activities from sensor data, in: *IJCAI*, 2015.
- [15] X. Chang, et al., Semantic pooling for complex event analysis in untrimmed videos, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (8) (2017) 1617–1632.
- [16] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: *Proceedings of the 15th ACM International Conference on Multimedia*, ACM, 2007.
- [17] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: *BMVC 2008-19th British Machine Vision Conference*, British Machine Vision Association, 2008.
- [18] H. Wang, et al., Action recognition by dense trajectories, in: *Computer Vision and Pattern Recognition (CVPR)*, 2011 IEEE Conference on, IEEE, 2011.
- [19] H. Wang, C. Schmid, Action recognition with improved trajectories, in: *Computer Vision (ICCV)*, 2013 IEEE International Conference on, IEEE, 2013.
- [20] H. Wang, et al., Evaluation of local spatio-temporal features for action recognition, in: *BMVC 2009-British Machine Vision Conference*, BMVA Press, 2009.
- [21] X. Peng, et al., Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, *Comput. Vis. Image Underst.* 150 (2016) 109–125.
- [22] L. Liu, L. Wang, X. Liu, In defense of soft-assignment coding, in: *Computer Vision (ICCV)*, 2011 IEEE International Conference on, IEEE, 2011.
- [23] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: *European Conference on Computer Vision*, Springer, 2010.
- [24] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: *European Conference on Computer Vision*, Springer, 2006.
- [25] F. Shi, R. Laganieri, E. Petriu, Gradient boundary histograms for action recognition, in: *Applications of Computer Vision (WACV)*, 2015 IEEE Winter Conference on, IEEE, 2015.
- [26] C. Caetano, J.A. dos Santos, W.R. Schwartz, Optical Flow Co-occurrence Matrices: A novel spatiotemporal feature descriptor, in: *Pattern Recognition (ICPR)*, 2016 23rd International Conference on, IEEE, 2016.
- [27] K. Muhammad, J. Ahmad, S.W. Baik, Early fire detection using convolutional neural networks during surveillance for effective disaster management, *Neurocomputing* (2017).
- [28] A. Karpathy, et al., Large-scale video classification with convolutional neural networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [29] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [30] E. Park, et al., Combining multiple sources of knowledge in deep cnns for action recognition, in: *Applications of Computer Vision (WACV)*, 2016 IEEE Winter Conference on, IEEE, 2016.
- [31] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: *Advances in Neural Information Processing Systems*, 2014.
- [32] S. Ji, et al., 3D convolutional neural networks for human action recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (1) (2013) 221–231.
- [33] S. Luo, et al., Real-time action recognition in surveillance videos using ConvNets, in: *International Conference on Neural Information Processing*, Springer, 2016.
- [34] J. Ahmad, et al., Efficient conversion of deep features to compact binary codes using fourier decomposition for multimedia big data, *IEEE Trans. Ind. Inform.* 14 (7) (2018) 3205–3215.
- [35] K. Muhammad, et al., Convolutional neural networks based fire detection in surveillance videos, *IEEE Access* 6 (2018) 18174–18183.
- [36] K. Muhammad, T. Hussain, S.W. Baik, Efficient CNN based summarization of surveillance videos for resource-constrained devices, *Pattern Recognit. Lett.* (2018).
- [37] C. Wang, et al., Image captioning with deep bidirectional LSTMs, in: *Proceedings of the 2016 ACM on Multimedia Conference*, ACM, 2016.
- [38] J. Deng, et al., Imagenet: A large-scale hierarchical image database, in: *Computer Vision and Pattern Recognition*, 2009. *CVPR 2009. IEEE Conference on, Ieee*, 2009.
- [39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [40] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Advances in Neural Information Processing Systems*, 2012.
- [41] P. Sermanet, et al., Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013. *arXiv preprint arXiv:1312.6229*.
- [42] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: *European Conference on Computer Vision*, Springer, 2014.
- [43] K. Muhammad, et al., Efficient deep CNN-based fire detection and localization in video surveillance applications, *IEEE Trans. Syst. Man Cybern.: Syst.* (99) (2018) 1–16.
- [44] I.N. Da Silva, et al., *Artificial Neural Networks*, Springer, 2017.
- [45] Q. Zhang, et al., An efficient deep learning model to predict cloud workload for industry informatics, *IEEE Trans. Ind. Inf.* (2018).
- [46] C. Hong, et al., Multimodal deep autoencoder for human pose recovery, *IEEE Trans. Image Process.* 24 (12) (2015) 5659–5670.
- [47] Z. Nazari, D. Kang, Density based support vector machines for classification, *IJARAI Int. J. Adv. Res. Artif. Intell.* (2015) 4–4.
- [48] R. Li, et al., Multi-objective optimization for rebalancing virtual machine placement, *Future Gener. Comput. Syst.* (2017).
- [49] I. Laptev, et al., Learning realistic human actions from movies, in: *Computer Vision and Pattern Recognition*, 2008. *CVPR 2008. IEEE Conference on, Ieee*, 2008.
- [50] C.A. Caetano, et al., Activity recognition based on a magnitude-orientation stream network, in: *Graphics, Patterns and Images (SIBGRAPI)*, 2017 30th SIBGRAPI Conference on, IEEE, 2017.
- [51] Z. Cai, et al., Multi-view super vector for action recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [52] N. Srivastava, E. Mansimov, R. Salakhudinov, Unsupervised learning of video representations using lstms, in: *International Conference on Machine Learning*, 2015.
- [53] A.-A. Liu, et al., Hierarchical clustering multi-task learning for joint human action grouping and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (1) (2017) 102–114.
- [54] W. Zhu, et al., A key volume mining deep framework for action recognition, in: *Computer Vision and Pattern Recognition (CVPR)*, 2016 IEEE Conference on, IEEE, 2016.
- [55] L. Sun, et al., Human action recognition using factorized spatio-temporal convolutional networks, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015.
- [56] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional Two-Stream Network Fusion for Video Action Recognition, 2016.
- [57] S. Ramasinghe, R. Rodrigo, Action recognition by single stream convolutional neural networks: an approach using combined motion and static information, in: *Pattern Recognition (ACPR)*, 2015 3rd IAPR Asian Conference on, IEEE, 2015.
- [58] K.K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, *Mach. Vis. Appl.* 24 (5) (2013) 971–981.
- [59] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: *Computer Vision and Pattern Recognition*, 2009. *CVPR 2009. IEEE Conference on, Ieee*, 2009.
- [60] H. Kuehne, et al., HMDB: a large video database for human motion recognition, in: *Computer Vision (ICCV)*, 2011 IEEE International Conference on, IEEE, 2011.



Amin Ullah received the bachelor's degree in computer science from the Islamia College Peshawar, Peshawar, Pakistan. He is currently pursuing the M.S. leading to the Ph.D. degree in digital contents with the Intelligent Media Laboratory, Sejong University, South Korea. His research interests include human action and activity recognition, sequence learning, image and video analysis, and deep learning for multimedia understanding.



Khan Muhammad received the bachelor's degree in computer science from Islamia College Peshawar, Peshawar, Pakistan, in 2014, with a focus on information security. He is currently an assistant professor in the Department of Software, Sejong University, South Korea. He has authored over 50 papers in peer-reviewed international journals and conferences, such as the *IEEE Transactions on Industrial Informatics*, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *IEEE Transactions on Industrial Electronics*, *Future Generation Computer Systems*, *Neurocomputing*, *IEEE Internet of Things Journal*, *PLoS One*, *IEEE ACCESS*, *Journal of Medical Systems*, *Biomedical Signal Processing and Control*, *Multimedia Tools and Applications*, *Pervasive and Mobile Computing*, *SpringerPlus*, *KSII Transactions on Internet and Information Systems*,

MITA 2015, PlatCon 2016, FIT 2016, Platcon-17, ICNGC-2017, and ICNGC-2018. His current research interests include image and video processing, information security, image and video steganography, video summarization, diagnostic hysteroscopy, wireless capsule endoscopy, computer vision, deep learning, and video surveillance. Mr. Muhammad is an Active Reviewer of over 30 reputed journals and is involved in editing of several special issues.



Ijaz Ul Haq received the bachelor's degree in computer science from the Islamia College Peshawar, Peshawar, Pakistan. He is currently pursuing the M.S. leading to the Ph.D. degree with the Intelligent Media Laboratory, Sejong University, South Korea. His research interests include movies summarization, image and video analysis, image hashing, steganography, and deep learning for multimedia understanding.



Sung Wook Baik received the B.S. degree in computer science from Seoul National University, Seoul, South 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, Fairfax, VA, USA, in 1999. He was with 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, 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, Sejong University. His research interests include computer vision, multimedia, pattern recognition, machine learning, data mining, virtual reality, and computer games. He served as a Professional Reviewer for several well-reputed journals, such as the IEEE Communication Magazine, Sensors, Information Fusion, Information Sciences, the IEEE TIP, MBEC, MTAP, SIVP, and JVICI.