Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM

Amin Ullah, *Student Member, IEEE*, Khan Muhammad, *Member, IEEE*, Javier Del Ser, *Senior Member, IEEE*, Sung Wook Baik, *Member, IEEE*, and Victor Hugo C. de Albuquerque, *Member, IEEE*

***Abstract*—Nowadays digital surveillance systems are universally installed for continuously collecting enormous amounts of data, thereby requiring human monitoring for the identification of different activities and events. Smarter surveillance is the need of this era through which normal and abnormal activities can be automatically identified using artificial intelligence and computer vision technology. In this paper, we propose a framework for activity recognition in surveillance videos captured over industrial systems. The continuous surveillance video stream is first divided into important shots, where shots are selected using the proposed convolutional neural network (CNN) based human saliency features. Next, temporal features of an activity in the sequence of frames are extracted by utilizing the convolutional layers of a FlowNet2 CNN model. Finally, a multilayer long short-term memory is presented for learning long-term sequences in the temporal optical flow features for activity recognition. Experiments[1] are conducted using different benchmark action and activity recognition datasets, and the results reveal the effectiveness of the proposed method for activity recognition in industrial settings compared with state-of-the-art methods.**

***Index Terms*—Activity recognition, artificial intelligence, convolutional neural network (CNN), deep learning, industrial systems, long short-term memory (LSTM), surveillance applications.**

Manuscript received August 2, 2018; revised October 13, 2018; accepted October 28, 2018. Date of publication November 22, 2018; date of current version July 31, 2019. This work was supported by the National Research Foundation of Korea funded by the Korean government (MSIP) under Grant 2016R1A2B4011712. *(Corresponding author: Sung Wook Baik.)*

A. Ullah and S. W. Baik are with the Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, South Korea (e-mail: aminullah@ieee.org; sbaik@sejong.ac.kr).

K. Muhammad is with Department of Software, Sejong University, Seoul 143-747, South Korea (e-mail: khan.muhammad@ieee.org).

J. Del Ser is with the TECNALIA, 48160 Derio, Bizkaia, Spain, and also with the University of the Basque Country (UPV/EHU) and Basque Center for Applied Mathematics (BCAM), Spain (e-mail: javier.delser@tecnalia.com).

V. H. C. de Albuquerque is with theLaboratory of Bioinformatics, University of Fortaleza, Fortaleza 60811-905, Brazil (e-mail: victor.albuquerque@unifor.br).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2018.2881943

I. Introduction

HUMAN activities and events analysis in visual surveillance are significantly beneficial and have recently evolved in the industrial sector, where hundreds of workers on jobs need automatic monitoring. Human activity recognition is a demand of numerous industrial applications, such as video summarization, smart surveillance and monitoring systems, virtual reality, robotics, medical diagnostic, elderly healthcare, and content-based video retrieval [1], [2]. Surveillance cameras are running in a 24/7 fashion. However, in different environments shots of interest occur rarely, like activity in front of an ATM machine, which does not happen continuously. Therefore, for such scenarios, an efficient shot segmentation is required for effective activity analysis. Shot segmentation is an important step of video processing, which helps us to analyze only vital contents in the stream for which several attempts have been made [3]. For instance, Mehrnaz *et al*. [4] presented a hidden-to-observable Markov model for shot boundary detection for automated soccer video analysis and classification. In another technique, it is used for background extraction using the temporal and spatial relationship in the input sequence. Similarly, Baohan *et al*. [5] and Irfan *et al*. [6] extracted shots to generate a diverse video summary. The mentioned existing methods are generic and not applicable to activity recognition in industrial settings. It is evident from our experiments that state-of-the-art techniques segmented shots, considering all sort of visual contents. However, video segmentation for activity recognition requires the presence of humans in each shot. Thus, for better activity recognition, we considered only human saliency features to perform shot segmentation that suppresses unessential information. Furthermore, processing only important shots minimizes the execution time, avoids extra computations, and improves the accuracy of activity recognition.

Complex real-world activity representation in visual data is a challenging task because of the viewpoint, pose, and scale of an actor. Currently, many researchers from artificial intelligence and computer vision society are working on the understanding of human activities and behaviors using traditional handcrafted features and neural network (NN) based approaches. For

[1]https://github.com/Aminullah6264/Activity_Rec_ML-LSTM

example, the earlier work [7] presented a hierarchical rank pooling that encodes sequence information of a video for activity recognition. It is a feed-forward NN of nonlinear operation and rank pooling followed by a Softmax classifier. Similarly, Sekma *et al*. [8] modified the traditional Fisher vector (FV) encoding method to propose a multilayer FV based on trajectory descriptors. They considered an advanced representation of geometric relationships between trajectories using three connected layers and local spatial pooling. Carlos *et al*. [1] combined different modalities for event analysis, including action recognition, object recognition, and knowledge-driven event recognition. In another approach by Xin *et al*. [9], long-term actions are recognized by utilizing recurrent-convolutional hybrid networks. They target the problem of spatial and temporal variations and interclass diversities through static and dynamic sequential features. In [10], Zhu *et al*. have jointly modeled different activities in a scene using motion information and contextual features by arguing that activities have a close relation with context. First, actions are detected and segmented followed by a two-layered conditional random field to recognize activities from the segmented patterns and contextual information. Recently, Jingyi *et al*. [11] presented the factorized action-scene network (FASNet) that fused three-dimensional (3-D) and two-dimensional (2-D) CNN networks that can effectively encode temporal action representations, followed by two loss layers to achieve specific action recognition tasks.

Besides CNNs, recurrent NN (RNN) based approaches are also presented by many researchers for action and activity recognition. An action classification approach in soccer videos using long short-term memory (LSTM) is proposed in [12]. First, they extracted a set of features at each time step which describes the visual content and dominant motion through Bag-of-Words (BoW) and interest point descriptors (IPD), respectively. They trained an LSTM network for learning video sequences using the temporal BoW and IPD features. Yue *et al*. [13] evaluated two CNN architectures to combine color and optical flow information, followed by LSTM network for action recognition. Ibrahim *et al*. [14] presented a two-stage temporal model to recognize group activities. They designed an LSTM model to represent action of an individual in a sequence of frames, whereas the second LSTM network aggregated individual level representation in the scene to understand the entire activity. Biswas *et al*. [15] proposed a structural RNN (SRNN) for activity recognition of underlining group in a shot. SRNN is a sequence of interconnected RNNs, which analyze the actions of humans, interactions between them, and their group activity. A recent approach introduced a CNN-based deep bidirectional LSTM [16] (DB-LSTM) for action recognition. They have used the AlexNet CNN model for frame-level feature extraction and DB-LSTM for sequence learning. However, CNN followed by DB-LSTM proves to be computationally very expensive and inefficient for activity recognition in industrial setup due to its designed pipeline consisting of huge convolutional and LSTM layers.

The current literature of human activity recognition indicates some popular methods, including BoW, spatio-temporal features [9], IPDs [12], optical flow trajectories [8], and CNN-based learned features [11]. The major issues in activity recognition are camera motion, similarity in visual contents, scattered background, viewpoint variations, and different lighting changes, which cannot be addressed with the aforementioned existing methods. The task of activity recognition becomes more challenging in industrial settings where the video stream is continuous and shots are inherently long. To address these problems, we propose a deep-learning-based framework for activity recognition for industrial surveillance applications with the following major contributions.

1) We propose an activity recognition framework for industrial systems, where human saliency is detected in the surveillance stream by investigating the convolutional layers of a pretrained CNN model MobileNet [17]. We have trained feature maps of CNN model on Institut National de Recherche en Informatique et en Automatique (INRIA) person dataset [18], which learns to select only those salient regions that are activated for persons in a video frame. Those regions are used for salient feature extraction and shot segmentation, resulting in representative shots suitable for industrial video stream analysis and activity recognition.
2) Optical flow is a highly preferable source of motion estimation in video sequences, and motion features are the key features for activity recognition; therefore, our framework utilized a CNN-based optical flow model FlowNet2 [19] for temporal features extraction. Our mechanism employs a new global average pooling (GAP) layer after the last convolutional layer for feature representation, because the convolutional features are local and have the capability to capture frame-to-frame variation in activity. This makes our framework more suitable for activity recognition in industrial settings.
3) Literature shows that LSTM outperformed other learning methods in learning time series data. Therefore, we present a multilayer LSTM for activity recognition using the temporal optical flow features. The multilayer structures and 256 cell size of LSTM empower our model to learn long-term activity sequences in video stream. We believe that our proposal is a suitable candidate for activity recognition in regular surveillance in general and industrial environment in particular.

We structure the remainder of this paper as follows. Section II covers all the technical details of our system, Section III provides the experimental results, evaluations, and discussion, and Section IV summarizes the work with key findings and suggests future research directions.

II. Proposed Activity Recognition Framework

In this section, the main mechanisms of the proposed approach are discussed in detail. The proposed approach is divided into three main steps with details given in Fig. 1 and mathematically presented in Algorithm 2. First, important shots are segmented from the continuous video stream using human saliency features. Second, frames of the segmented shots are propagated forward to the optical flow CNN model FlowNet2 to extract temporal features from the intermediate layer using the proposed procedure. Finally, the activity in the segmented shot is recognized using the proposed trained multilayer LSTM network.

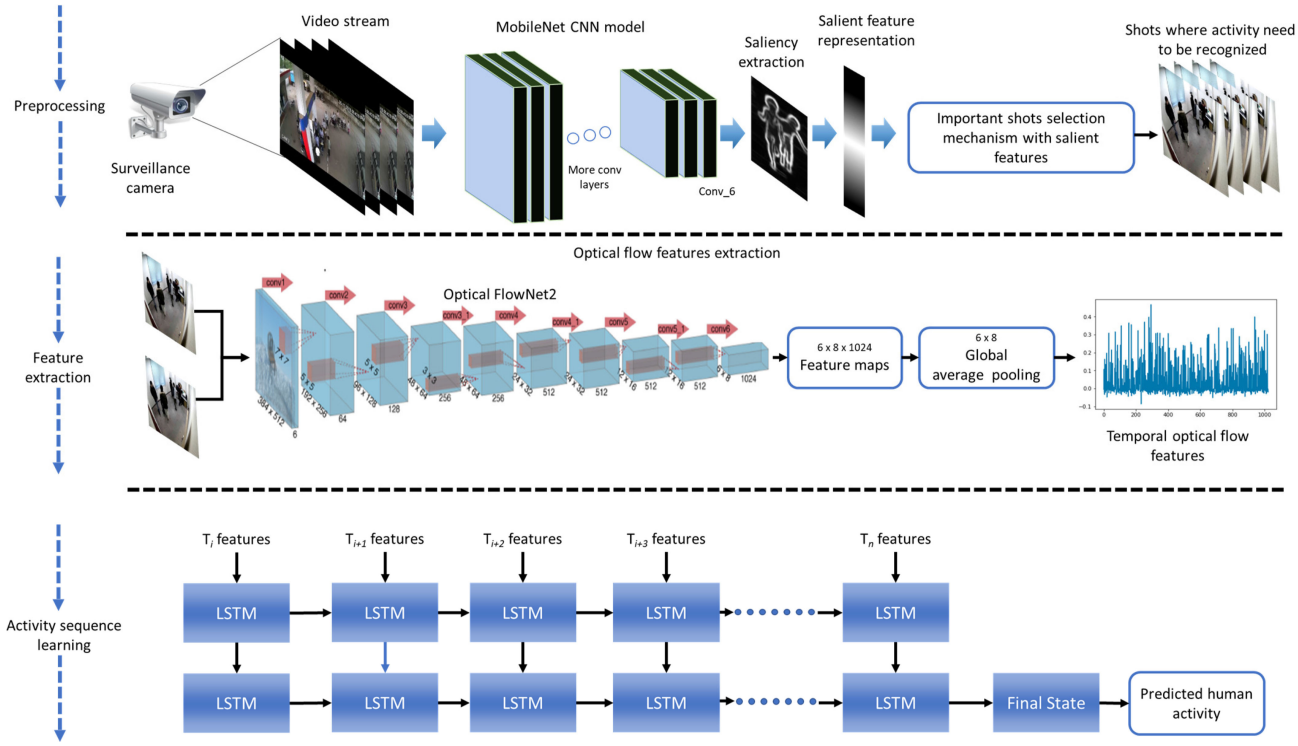

Fig. 1. Proposed framework for activity recognition in industrial systems. In the first stage, the surveillance stream is preprocessed using our CNN-based salient feature extraction mechanism for shots segmentation. The second stage extracts CNN-based temporal optical flow features from the sequence of frames. Finally, multilayer LSTM is presented for sequence learning and recognition of activities.

A. Shot Segmentation

Shot segmentation is an important step within a surveillance video processing pipeline. We have analyzed the existence of humans in the video stream using the convolutional features of the MobileNet [17] CNN model that is pretrained on largescale ImageNet dataset of millions of images. Convolutional layers are backbone of feature learning process of CNN. These layers have the essential capacity to learn complex local patterns from the visual data where the fully connected layers can learn the global high-level semantic representation of the image. Convolutional layers consist of different feature maps acquired from the number of kernels applied at learning stage. MobileNet [17] is a highly inspired neurological CNN model which applies small-sized kernels to images, making it efficient in terms of time complexity and can learn tiny hidden patterns over visual data.

Recently, many studies [20]–[23] have investigated the mentioned feature maps for object detection, image retrieval, and fire detection and localization problem. In this paper, we investigate it for specific object (human) salient features extraction. MobileNet [17] has a huge pipeline of convolutional layers, wherein we have utilized the output of "conv5" layer having $14 \times 14 \times 512$ dimensions. The "conv5" layer has 512 feature maps, where the activation in most maps is not representing the human body structure. Therefore, we have utilized only those features that are highly activated for humans and avoid irrelevant features. In Fig. 2(b), highly activated features expose the most salient regions (human) in images while ignoring the background information. The procedure for salient feature maps selection is given in Algorithm 1.

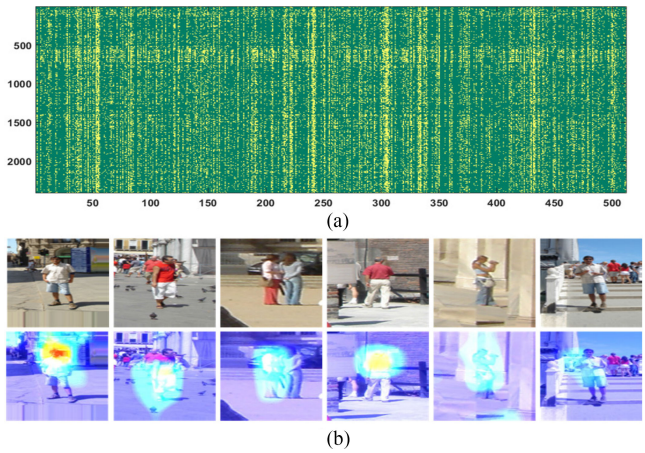

Fig. 2. (a) SEI where the green color represents those feature maps that give high saliency and the yellow color means the particular feature maps did not respond to the salient object. (b) Highly activated regions selected by our proposed mechanism for salient features extraction aimed at shot segmentation of surveillance stream.

The proposed mechanism for salient feature selection is trained on INRIA person dataset [18], which consists of more than 2000 images of human collected using surveillance cameras of both indoor and outdoor scenes. First, we have extracted $14 \times 14 \times 512$ feature maps from the conv5 layer of MobileNet for all images in the database. Second, global mean activation value for each feature map is calculated for all training set. Afterward, the feature maps that respond to humans in the dataset are marked as 1 [green colored in the saliency exploitation index (SEI) as

Algorithm 1: Salient Feature Map Selection for Shot Segmentation.

Input: Images Dataset (ID)
Preparation:
1: Load MobileNet-V2 CNN Model
2: F_N Feature Maps in 'conv5'
3: Saliency Exploitation Index (SEI) of size (ID) $\times$ F_N.
Steps:
1: **for each** image ID_i **in** ID
a: Feed ID_i to MobileNet-V2-CNN
b: Extract $h \times w \times F_N$ feature maps from layer "conv5"
c: Compute global mean *GMi* of each feature map *Fi* to obtain F_N values
d: Locate feature maps whose $GMi = 0$.
e: Mark their indices with 1 in the SEI for ID_i
end for
2: Histogram of null values $HIST\_NV_i$ for each *Fi*
3: Percentages of null values *P_HIST_NVi* for each *Fi*
Return all *Fi* as *SF* whose *P_HIST_NVi* $<$ Threshold
Output: Selected Salient Feature Maps (S_F)

shown in Fig. 2(a)]. The feature maps that are not activated for human regions are represented as 0 [yellow colored in the SEI as shown in Fig. 2(a)]. This procedure discards the convolutional kernels that generate zero activations for majority of the training images. Next, the histogram and percentage of all null values are calculated in the SEI map. Finally, we have set a threshold value *t* and the feature maps having null values percentage less than *t* are selected as salient feature maps. Activations of the selected feature maps are shown in Fig. 2(b), which exhibit high attention to the human regions in the images. The proposed algorithm intelligently selects highly activated feature maps and discards the ones that are not activated for training images.

1) Feature Extraction Using Salient Feature Maps: The frames of continuous video stream are propagated forward to MobileNet for salient features extraction, where only those feature maps are extracted from layer "conv5" which are selected using training phase of Algorithm 1. The key role of Algorithm 1 is to select only those feature maps that are highly activated and salient for human regions. In each selected feature map, we have calculated a global mean that represents the activation of human in each feature map. After the extraction of feature vectors from two consecutive frames of the video stream, we have calculated the Euclidean distance between both feature vectors. If the distance value surpasses a particular threshold *t*, it indicates distinction of current shot from the previous one. The threshold is set after analyzing distance values given by different images containing relevant and irrelevant visual data. The effectiveness of our approach of selecting shots based on saliency information can be verified from Fig. 3. A sample video is segmented into shots using our approach and another CNN-based method [24] and the results are given in Fig. 3. The approach [23] selects shots based on all kind of visual contents without giving any preference to humans while our approach focuses on shots having humans.

Fig. 3. Shot segmentation results for a sample video. (a) Representative frames of eight shots obtained by the existing CNN-based approach [24]. (b) Results of our proposed shot segmentation approach.

Algorithm 2: Activity Recognition.

Input: Surveillance video stream
Preparation:
1: MobileNet-V2 CNN model
2: Saliency exploitation index (selected feature maps)
3: Optical flow CNN model (FlowNet2)
4: Trained multilayer LSTM
Steps:
1: **while** (video stream)
2: Forward propagate frame to MobileNet-V2 CNN
3: Extract conv5 selected feature maps (SEI $\leftarrow$ **Algorithm 1**)
4: Calculate distance between two consecutive frames to select shot of interest
5: **if** (Shot of interest)
a: Feed frame I *and* frame I+1 to FlowNet2-CNN
b: Extract $h \times w \times F_N$ feature maps from layer "conv6"
c: Compute global average pooling (GAP) GP_i of each feature map *Fi* to obtain *Fv* feature vector
d: Feature vector *Fv* is input to the trained LSTM at timestep 't'
6: Predict activity from $\leftarrow$ **Trained LSTM**
7: Display predicted activity with confidence score
end if
end while
Output: Labeled activity along with confidence score

B. Temporal Optical Flow Features Extraction

Human activity is the motion of body parts changing in consecutive video frames. Optical flow is the most powerful motion detection approach in sequence of images, which is used in a wide range of domains, including moving picture experts group compression, video restoration, vehicle navigation, video indexing and retrieval, flow of blood cellular tissues, and biomedical data analysis [25], [26]. We exploited it for human activity recognition in real-world data, which consist of flow of human body parts. However, small-displacement optical flow detection in real-world data is still a very challenging task in computer vision. To overcome this problem, we utilized

CNN-based optical flow method known as FlowNet2 [19]. It is trained for slight motion displacement data, in which the displacement histogram of training data is similar to UCF101 [27] action dataset, which consists of real-world human action videos. Therefore, we argue that feature extraction using FlowNet2 convolutional layers can effectively represent human activities in sequence of frames as it can capture large- and small-displacement optical flow, making it suitable for industrial settings.

FlowNet2 is originally designed for end-to-end optical flow detection, which is trained on consecutive image pairs and its labeled flow. This clashes with traditional CNN models, which forward propagate one image and learn its hidden patterns. In FlowNet2, the initial layers process both images separately using convolutional kernels, which extract the semantic visual representation of an image. In the middle of the network, convolutional features of both images are combined using "correlation layer," which performs multiplicative patch comparisons between two feature maps followed by a pipeline of convolutional layers. It is proven in many recent studies that intermediate layers of CNN architecture are very powerful, as they can represent the local motion patterns in convolutional layers. The initial layers of CNN model analyze only a small neighborhood that contains primitive features, whereas deeper layers have comparatively wide receptive fields and can capture high-level motion semantics [28].

In this paper, we have explored the final convolutional layer of FlowNet2 model, which is pretrained for flow estimation. This layer represents the actual motion in consecutive frames, which is further deconvoluted to represent the detected flow. Two consecutive frames are fed to the pretrained FlowNet2 CNN model. Feature maps from final convolutional layers ($h \times w \times d$) are extracted. For each d feature map, GAP is applied to get a dense motion feature, which is global representation of features maps. GAP is already used by several recent state-of-the-art CNN models [29], [30] for classification and object detection tasks due to its effective visual data representation. GAP gives one value for the entire feature map in which the kernel size is $h \times w$. GAP layer is used to reduce the spatial dimensions of a 3-D feature map. However, GAP layers perform a more extreme sort of dimensionality reduction, in which a feature map having dimensions $h \times w \times d$ is reduced in size to have d dimensions. To get a d-dimensional temporal optical flow feature vector, all GAP results are combined layerwise, representing the final temporal optical flow feature for two connected frames.

C. Learning Activities via Multilayer LSTM

LSTM is the extended form of RNN, which has the ability to find temporal-sequential patterns in time series data. LSTM is the solution of the vanishing gradient problem of simple RNN, which is not able to learn long-term sequences and loss the effect of initial dependencies in the sequence. LSTM involves input, forget, and output gates that helps to learn the long-term sequence information. Naïve RNN takes an input at every time step; however, the LSTM unit decides whether to take the input by using a sigmoid layer for each gate to be open or closed. As shown in (1)–(3) i_t, f_t, and o_t are input, forget, and output gates, respectively. In (4), g is the recurrent unit, which is also called standard RNN, that is calculated from the input at the current time x_t and hidden state of the previous time step S_{t-1}. Equation (6) calculates the S_t current hidden state of the LSTM at time "t" through tanh activation and (5) memory cell c_t. The memory cell decides the contribution of the previous time step and current input to calculate hidden state S_t. The final state of the LSTM network represented in (8) as S_{tN} is basically the final representation of the whole sequence which is passed through Softmax classifier for activity prediction:

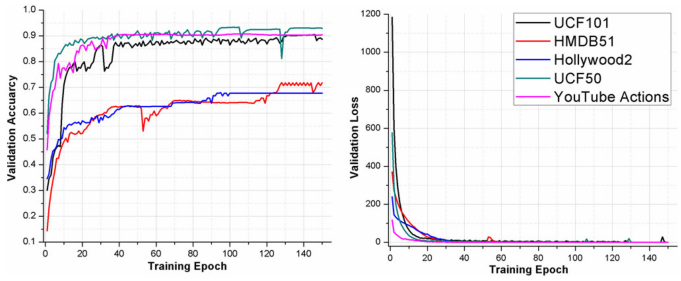

Fig. 4. Learning progress of the proposed multilayer LSTM for 150 training epochs.

$$i_t = \sigma\left(\left(x_t + S_{t-1}\right) W^i + b_i\right) \quad (1)$$

$$f_t = \sigma\left(\left(x_t + S_{t-1}\right) W^f + b_f\right) \quad (2)$$

$$o_t = \sigma\left(\left(x_t + S_{t-1}\right) W^o + b_o\right) \quad (3)$$

$$g = \tanh\left(\left(x_t + S_{t-1}\right) W^g + b_g\right) \quad (4)$$

$$c_t = c_{t-1}.f_t + g.i_t \quad (5)$$

$$s_t = \tanh\left(c_t\right).o_t \quad (6)$$

$$s_t^{\mathrm{l}} = \tanh\left(c_t^{\mathrm{l}}\right).o_t^{\mathrm{l}} \quad (7)$$

$$\text{predictions} = \text{Softmax}(s_{tN}). \quad (8)$$

Deep NNs (DNNs) have been widely used in computer vision society and surpasses in different domains, including image classification, image retrieval, and object detection [31]–[34]. In the proposed framework, we have applied DNN style LSTM by applying multilayer LSTM for activity recognition. In multilayer LSTM the state of layer "l" gets input from the previous layer and the previous state of the same layer (7). In a typical DNN, the activation values of the neurons often have hundreds of dimensions. Thus, those activations are capable of learning patterns in huge data. Therefore, we argue that by stacking multiple layers of LSTM, its ability to learn long-term sequences is dramatically increased. We have extracted temporal optical flow features of 1024 dimensions from two consecutive frames, which is input at time step t to multilayer LSTM. We forward propagate temporal features of 1-s video to LSTM in 15-time steps for learning activity sequence. The proposed LSTM contains 256 cells, which is trained up to 500 epochs along with 20% validation samples. The learning rate for cost minimization is kept to 0.001 using stochastic optimization. It can be observed from Fig. 4 that the proposed LSTM architecture has better ability for learning complex motion patterns and

TABLE I
TIME TAKEN BY EACH DATASET FOR FEATURE EXTRACTION AND MULTILAYER LSTM PARAMETER LEARNING

Dataset	Feature Extraction Time (hours)	Training Time (minutes)
UCF101	25.3	47.75
HMDB51	13.7	25.6
Hollywood2	5.8	14.9
UCF50	10.2	20.6
YouTube Actions	5.3	10.3

TABLE II
COMPARISON USING RECOGNITION SCORE WITH STATE-OF-THE-ART METHODS FOR UCF101 ACTIVITIES DATASET

Method	Accuracy
LSTM with 30 frame unroll [13]	88.6%
RLSTM-g3 [42]	86.9%
Hierarchical clustering multi-task [43]	76.3%
Factorized spatio-temporal CNNs [44]	88.1%
DB-LSTM [16]	92.84%
Proposed approach	**94.45%**

reducing the validation loss without going into overfitting problem for activity recognition.

III. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, our proposed activity recognition approach is experimentally assessed using different recognition accuracy, metrics including confusion matrix, overall accuracy, and classwise accuracy graph. The proposed approach is evaluated through different benchmark datasets comprising UCF101 [27], UCF50 [35], HMDB51 [36], YouTube [37], and Hollywood actions [38]. The performance of our approach is compared with state-of-the-art activity recognition methods. The proposed system is implemented and assessed in Python 2.7 on Ubuntu16.04, Corei5-6600 setup with 16-GB RAM and equipped with a 12-GB GeForce-Titan-X graphics processing unit (GPU). The deep learning toolbox "Caffe" is used for temporal optical flow feature extraction using CNN model FlowNet2, whereas deep learning framework "Tensorflow" is used for multilayer LSTM implementation. Experiments for all five datasets have been conducted using a stratified 60% sample for training, 20% for validation, and 20% for testing [39]–[41]. Feature extraction and training time of each dataset is given in Table I. Validation loss and accuracies for 150 epochs are shown in Fig. 4. Results of each dataset and comparisons with state-of-the-art techniques are discussed in separate sections.

A. UCF101 Dataset

In the literature related to activity recognition, UCF101 [27] is considered as a very challenging dataset because its videos resemble real-world activities. The overall dataset has 101 classes with a total of 13 320 videos selected from YouTube. In every class there are 100–200 activity samples performed by different actors, which represent five major types of categories, namely:

1) human–object interaction;
2) human–human interaction;
3) body-motion only;
4) playing musical instruments;
5) sports activities.

On this dataset, the proposed approach is compared with five activity recognition methods, including a single-layer LSTM [13], regularizing long short term memory (RLSTM-g3) [42], hierarchical clustering multitask (HCMT) [43], factorized spatiotemporal convolutional network (FSTC) [44], and DB-LSTM [16]. The obtained results are shown in Table II, and classwise accuracy on UCF101 dataset is given in Fig. 5. LSTM [13] with 30 frames unroll reported 88.6% accuracy, whereas RLSTM-g3 [42], HCMT [43], and FSTC [44] reported 86.9%, 76.3%, and 88.1% accuracy, respectively. The proposed approach improved accuracy score by 1.61% from this best accuracy recently achieved by DB-LSTM [16].

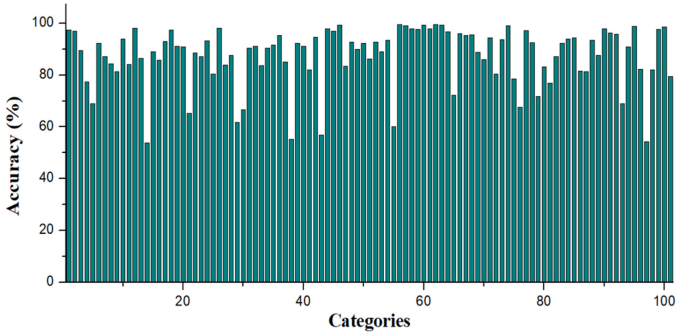

Fig. 5. Categorywise accuracy for the test set of UCF101 dataset for our activity recognition approach.

In Fig. 5, horizontal axis shows classes and the vertical axis represents percentage accuracies for the corresponding class for the test set of UCF101 dataset. From this figure, best accuracies are above 85%, approaching 100% for some categories. Only few accuracies are under 65% to 75%. Confusion matrix calculated from the test set of this dataset is visualized in Fig. 7(a), in which true positive intensities are brighter for all most all classes, showing the effectiveness of the proposed approach on UCF101 dataset.

B. HMDB51 Dataset

HMDB51 dataset is collected from various sources, mostly from movies and small proportion from public repositories, such as Google and YouTube videos. The dataset comprises of various types of activities related to general facial action with object manipulation, general body movements, and body movements for human interactions. It consists of 6766 sample videos divided into 51 different activity categories, and each of the categories contains more than 100 video clips. The duration of video clips for most of the activities is less than 5 s in this dataset. Using the HMDB51 dataset, the proposed method is compared with five activity recognition techniques, including RLSTM-g3 [42], HCMT [43], FSTC [44], A-RNN [9], and multilayer fisher vector (MLFV) [8].

The obtained results on the test set of HMDB51 dataset are reported in Table III. The categorywise accuracies are given in Fig. 6, and confusion matrix is shown in Fig. 7(b). For this dataset, RLSTM-g3 [42] reported 55.3% accuracy whereas

TABLE III
COMPARISON USING RECOGNITION SCORE WITH STATE-OF-THE-ART METHODS FOR HMDB51 ACTIVITIES DATASET

Method	Accuracy
RLSTM-g3 [42]	55.3%
Hierarchical clustering multi-task [43]	51.4%
Factorized spatio-temporal CNNs [44]	59.1%
Adaptive RNN-CNNs [9]	61.1%
Multi-layer fisher vector [8]	68.5%
Proposed Approach	**72.21%**

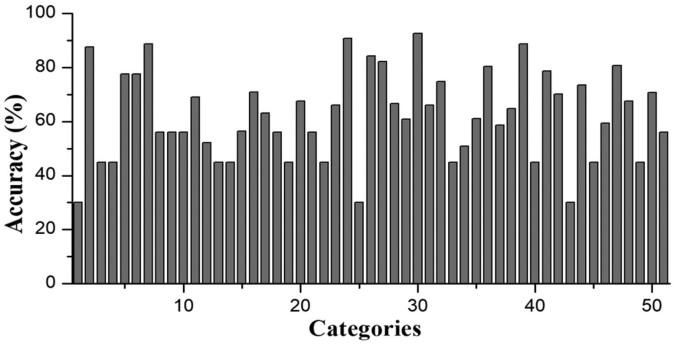

Fig. 6. Categorywise accuracy for the test set of HMDB51 dataset using the proposed activity recognition approach.

HCMT [43], FSTC [44], A-RNN [9], and MLFV [8] obtained 51.4%, 59.1%, 61.1%, and 68.5%, respectively. The proposed approach increases 3.5% accuracy score on this dataset. In this challenging dataset, our approach performed best. From Fig. 6 (categorywise accuracy) and Fig. 7(b) (confusion matrix) it can be noted that few category scores are above 90%, but most of them are within the range 70%–90%. The accuracy for all classes is greater than 45% except kicking football, picking, and swing baseball. It is due to the overlapping of visual contents and similarity of motion information with other classes. Therefore, predictions of these classes are confused with others.

C. Hollywood2 Dataset

The Hollywood dataset provides 12 activities of humans. This dataset is very comprehensive and considered as a benchmark in activity recognition literature. This repository is created from 69 Hollywood movies containing 810 video clips in Avi format. Using this dataset, our approach is compared with different state-of-the-art schemes, including visual attention [45], hierarchical rank pooling [7], improved trajectory [46], multiskip feature stacking (MSFS) [47], and FASNet + MIFS + support vector machine (SVM) [11]. The accuracies obtained on this dataset are given in Table IV, categorywise accuracies are given in Fig. 8, and confusion matrix is shown in Fig. 7(c). This dataset is one of the most challenging datasets for activity recognition evaluation, due to the fact that many overlapping activities are performed within each category. For example, an activity "Standing" is performed next to an eating table, which is, in turn, also represented in activity "Eating." For this dataset, the visual attention-based method achieved 43.9%, whereas hierarchical rank pooling [7], improved trajectory [46], MSFS [47], and FASNet + MIFS + SVM [11] reported 56.8%, 64.3%, 68.0%, and 78.1%, respectively. The proposed approach did not achieve higher accuracy than all methods. The technique with higher accuracy than our method is the fusion of 3-D and 2-D CNNs [11]. In this regard, we argue that it is not an efficient classification method in terms of complexity for activity recognition in industrial setup compared with our approach. Thus, our method is the best candidate for industrial systems, considering both accuracy and implementation feasibility.

D. UCF50 Dataset

UCF50 comprises a variety of human actions, yielding one of the most resorted datasets in action recognition literature. It has a total of 50 action classes, and there are certain categories where different groups have mutual features variations. For instance, the same action is performed, but only with a different viewpoint. Using UCF50 dataset, the proposed approach is compared with five activity recognition methods, including latent structural SVM [48], effective event models [49], motion trajectories [50], improved trajectory [46], and hierarchical clustering multitask [43]. The obtained results are shown in Table V, categorywise accuracies are given in Fig. 9, and the confusion matrix for the test set of UCF50 dataset is shown in Fig. 7(d).

It is evident from Table V that the proposed approach has achieved higher accuracy for this dataset when compared to latent structural SVM [48], effective event models [49], motion trajectories [50], improved trajectory [46] and hierarchical clustering multitask [43] which gives 84.77%, 86.01%, 89.4%, 91.2%, and 93.2% accuracy, respectively. Fig. 9 shows categorywise accuracies, where most of the categories reported more than 90% accuracy results, which elucidate the robustness of our approach for different kinds of activity recognition in real-world scenarios.

E. YouTube Dataset

The YouTube action dataset is challenging because of very diverse videos with dynamic and static camera. There are some sports videos and other videos gathered from YouTube. This dataset has 11 action classes of sports category, and videos are taken from 25 participants with four paradigms per action. The comparative results using this dataset with other methods are given in Table VI, categorywise accuracies are shown in Fig. 10, and the confusion matrix for the test set is shown in Fig. 7(e). This dataset is relatively not as challenging as the previous ones; however, the proposed approach improved results on this dataset. We have improved 2.7% accuracy on this dataset compared with recently published DB-LSTM [16] with 92.84% and single-stream CNN [51] with 93.1% accuracies. From Fig. 10, it can be seen that for each category the proposed approach achieved more than 90% activity recognition result.

F. Computational Complexity of the Proposed Approach

The results given in Fig. 11 demonstrate that using GPU the proposed approach is admittedly designed nearly real time for activity recognition. The pictorial representation shows that the

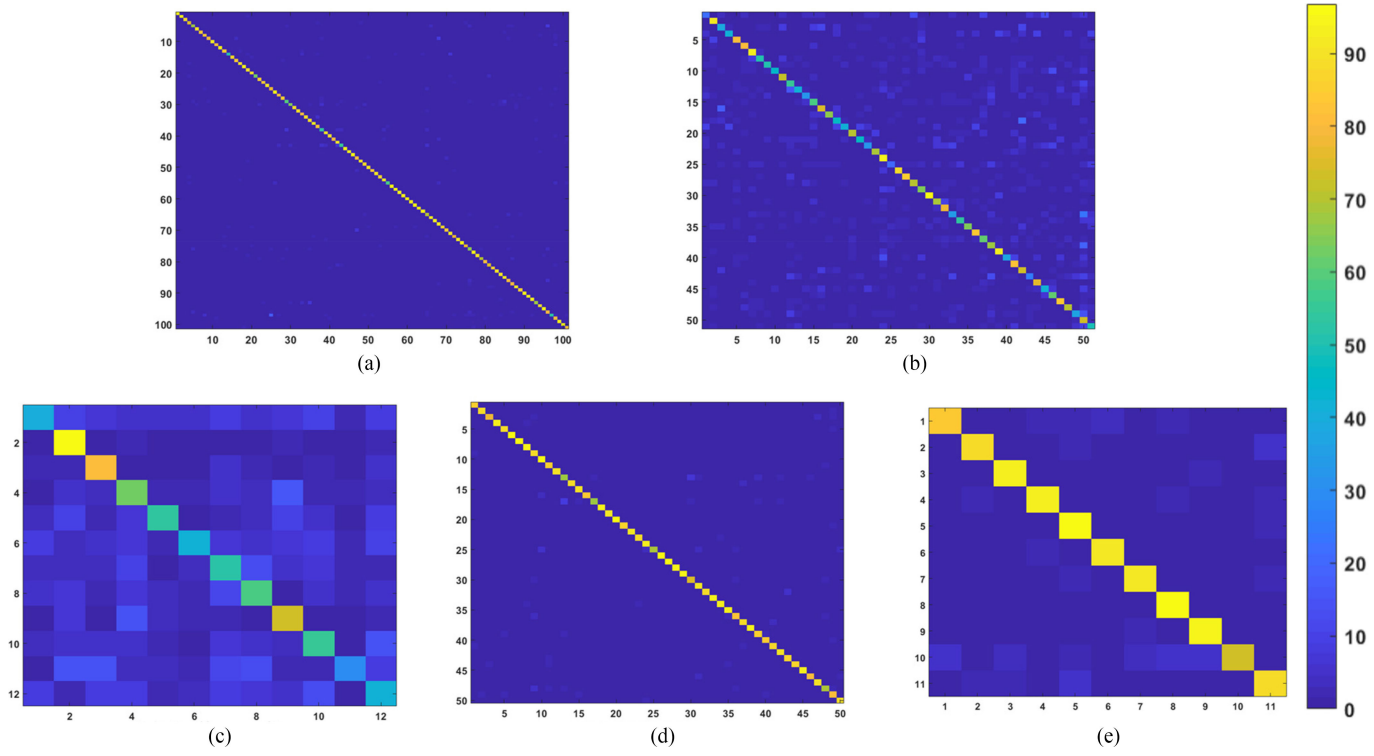

Fig. 7. Confusion matrices for testing five datasets used in the evaluation of our approach. (a) UCF101 dataset. (b) HMDB51 dataset. (c) Hollywood2 dataset. (d) UCF50 dataset. (e) YouTube dataset.

TABLE IV
COMPARISON USING RECOGNITION SCORE WITH STATE-OF-THE-ART METHODS FOR HOLLYWOOD2 ACTIVITIES DATASET

Method	Accuracy
Visual attention [45]	43.9%
Hierarchical rank pooling [7]	56.8%
Improved trajectory [46]	64.3%
Multi-skip feature stacking (MSFS) [47]	68.0%
FASNet + MIFS + SVM [11]	78.1%
Proposed Approach	**69.5%**

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS USING ACCURACY SCORE ON UCF50 DATASET

Method	Accuracy
Latent structural SVM [48]	84.77%
Effective event models [49]	86.01%
Motion trajectories [50]	89.4%
Improved trajectory [46]	91.2%
Hierarchical clustering multi-task [43]	93.2%
Proposed Approach	**94.9%**

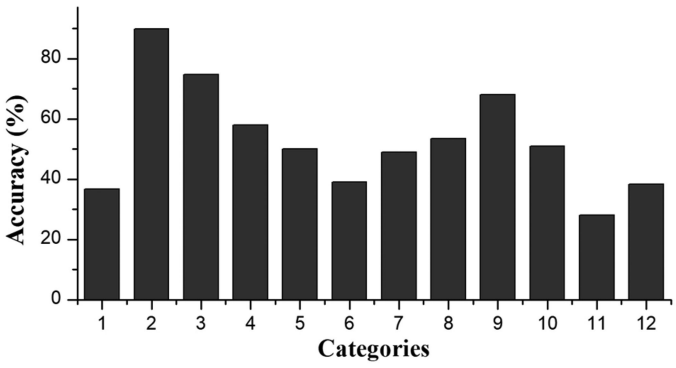

Fig. 8. Categorywise accuracy of Hollywood2 dataset for the proposed activity recognition approach.

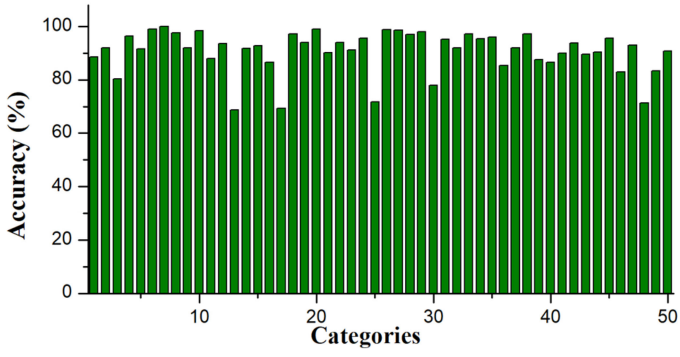

Fig. 9. Categorywise accuracy of UCF50 dataset for the proposed activity recognition approach.

process of salient feature extraction takes only 0.24 s, temporal features extraction takes 0.82 s, and recognition using multilayer LSTM consumes 0.008 s, respectively. Thus, the total time consumed shows that it is five times faster than a CPU-based setup. Therefore, the proposed approach is a better option to be used in any sort of industrial environment because of the fast computation of the proposed algorithm and higher accuracy. The

TABLE VI
COMPARISON USING ACCURACY SCORE WITH PREVIOUS METHODS FOR YOUTUBE ACTIVITIES DATASET

Method	Accuracy
Latent structural SVM [48]	85.97%
Hierarchical clustering multi-task [43]	89.7%
Discriminative representation [52]	91.6%
DB-LSTM [16]	92.84%
Single stream CNN [51]	93.1%
Proposed Approach	**95.8%**

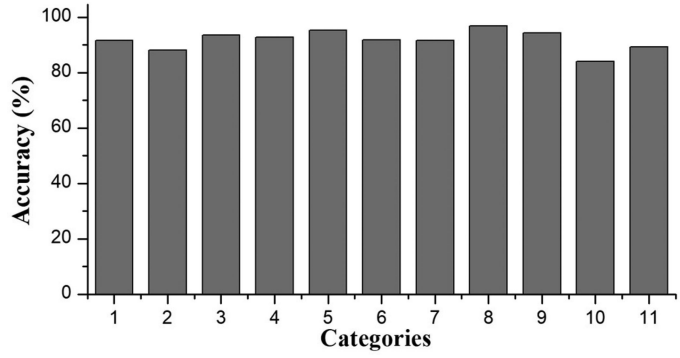

Fig. 10. Categorywise accuracy of YouTube actions dataset for the proposed activity recognition approach.

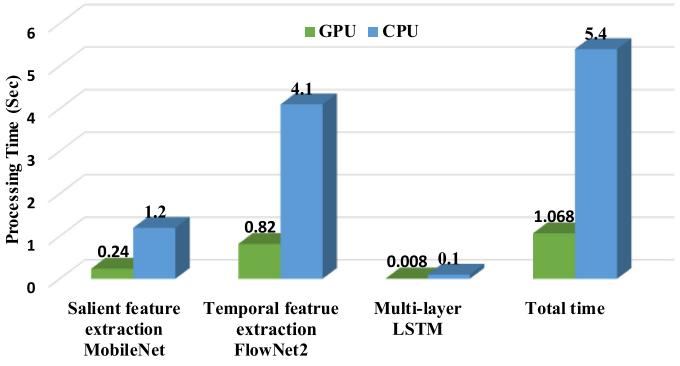

Fig. 11. Time complexity for processing 30 frames on CPU and GPU using the proposed approach for activity recognition in industrial environment.

above discussed tables and figures for five datasets verify that our system improved results on four datasets and the accuracy for all categories results is uniformly superior for most of the datasets.

IV. CONCLUSION

In this paper, we proposed an activity recognition approach for industrial surveillance systems. For activity analysis, we need only those parts from the surveillance stream where humans appear. To achieve this goal, first, important shots were captured from the video stream, where shots were selected using CNN-based human saliency features. Second, an activity in the selected shot was represented using temporal optical flow features by utilizing the convolutional layers of a FlowNet2 CNN model. Finally, a multilayer LSTM was presented for learning long-term sequences in the temporal optical flow features for activity recognition. Extensive experiments and comparisons have been conducted using different accuracy matrices for five benchmark activity recognition datasets, which verify the effectiveness of the proposed system. The time complexity of the system suits realistic processing requirements imposed by industrial setups nearly real-time video stream processing and high detection accuracies. This empirical finding proves the robustness of our approach for industrial surveillance applications.

The current approach performs well for activity recognition of a single person in the video. In the future, we will investigate individual and group activities for surveillance applications. Furthermore, we will also consider multiview data for action and activity recognition.

REFERENCES

[1] C. F. Crispim-Junior *et al.*, "Semantic event fusion of different visual modality concepts for activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1598–1611, Aug. 2016.
[2] M. Sajjad, A. Ullah, J. Ahmad, N. Abbas, S. Rho, and S. W. Baik, "Integrating salient colors with rotational invariant texture features for image representation in retrieval systems," *Multimedia Tools Appl.*, vol. 77, pp. 4769–4789, 2018.
[3] X. Liang and Z. Yan, "A survey on game theoretical methods in human–machine networks," *Future Gener. Comput. Syst.*, to be published, doi: 10.1016/j.future.2017.10.051.
[4] M. Fani, M. Yazdi, D. A. Clausi, and A. Wong, "Soccer video structure analysis by parallel feature fusion network and hidden-to-observable transferring Markov model," *IEEE Access*, vol. 5, pp. 27322–27336, 2017.
[5] B. Xu, X. Wang, and Y.-G. Jiang, "Fast summarization of user-generated videos: Exploiting semantic, emotional, and quality clues," *IEEE MultiMedia*, vol. 23, no. 3, pp. 23–33, Jul./Sep. 2016.
[6] I. Mehmood, M. Sajjad, S. Rho, and S. W. Baik, "Divide-and-conquer based summarization framework for extracting affective video content," *Neurocomputing*, vol. 174, pp. 393–403, 2016.
[7] B. Fernando, P. Anderson, M. Hutter, and S. Gould, "Discriminative hierarchical rank pooling for activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1924–1932.
[8] M. Sekma, M. Mejdoub, and C. B. Amar, "Human action recognition based on multilayer fisher vector encoding method," *Pattern Recognit. Lett.*, vol. 65, pp. 37–43, 2015.
[9] M. Xin, H. Zhang, H. Wang, M. Sun, and D. Yuan, "Arch: Adaptive recurrent-convolutional hybrid networks for long-term action recognition," *Neurocomputing*, vol. 178, pp. 87–102, 2016.
[10] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury, "Context-aware activity modeling using hierarchical conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1360–1372, Jul. 2015.
[11] J. Hou, X. Wu, Y. Sun, and Y. Jia, "Content-attention representation by factorized action-scene network for action recognition," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1537–1547, Jun. 2018.
[12] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Action classification in soccer videos with long short-term memory recurrent neural networks," in *Proc. Int. Conf. Artif. Neural Netw.*, 2010, pp. 154–159.
[13] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4694–4702.
[14] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1971–1980.
[15] S. Biswas and J. Gall, "Structural recurrent neural network (SRNN) for group activity analysis," in *Proc. 2018 IEEE Winter Conf. Appl. Comput. Vis.*, Lake Tahoe, NV, 2018, pp. 1625–1632.
[16] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.
[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
[19] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, p. 6.
[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," 2014, arXiv:1412.6856.
[21] J. Ahmad, K. Muhammad, S. Bakshi, and S. W. Baik, "Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets," *Future Gener. Comput. Syst.*, vol. 81, pp. 314–330, 2018.
[22] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30–42, 2018.
[23] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep CNN-based fire detection and localization in video surveillance applications," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2018.2830099.
[24] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient CNN based summarization of surveillance videos for resource-constrained devices," *Pattern Recognit. Lett.*, to be published, doi: 10.1016/j.patrec.2018.08.003.
[25] Y. Liu, L. Nie, L. Han, L. Zhang, and D. S. Rosenblum, "Action2Activity: Recognizing complex activities from sensor data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 1617–1623.
[26] I. Sobron, J. Del Ser, I. Eizmendi, and M. Velez, "A deep learning approach to device-free people counting from WiFi signals," in *Proc. Int. Symp. Intell. Distrib. Comput.*, 2018, pp. 275–286.
[27] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, arXiv:1212.0402.
[28] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4749–4757.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[31] A. Chadha and Y. Andreopoulos, "Voronoi-based compact image descriptors: Efficient region-of-interest retrieval with VLAD and deep-learning-based descriptors," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1596–1608, Jul. 2017.
[32] J. Ahmad, K. Muhammad, J. Lloret, and S. W. Baik, "Efficient conversion of deep features to compact binary codes using Fourier decomposition for multimedia Big Data," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3205–3215, Jul. 2018.
[33] J. L. Lobo, I. Laña, J. Del Ser, M. N. Bilbao, and N. Kasabov, "Evolving spiking neural networks for online learning over drifting data streams," *Neural Netw.*, vol. 108, pp. 1–19, 2018.
[34] Q. Zhang, L. T. Yang, Z. Yan, Z. Chen, and P. Li, "An efficient deep learning model to predict cloud workload for industry informatics," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3170–3178, Jul. 2018.
[35] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Mach. Vis. Appl.*, vol. 24, pp. 971–981, 2013.
[36] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis,* 2011, pp. 2556–2563.
[37] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,"' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2009, pp. 1996–2003.
[38] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2929–2936.
[39] H. K. Turesson, S. Ribeiro, D. R. Pereira, J. P. Papa, and V. H. C. de Albuquerque, "Machine learning algorithms for automatic classification of marmoset vocalizations," *PloS One*, vol. 11, 2016, Art. no. e0163041.
[40] S. Liu, L. Feng, J. Wu, G. Hou, and G. Han, "Concept drift detection for data stream learning based on angle optimized global embedding and principal component analysis in sensor networks," *Comput. Elect. Eng.*, vol. 58, pp. 327–336, 2017.
[41] R. J. Mstafa and K. M. Elleithy, "A new video steganography algorithm based on the multiple object tracking and Hamming codes," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl.*, 2015, pp. 335–340.
[42] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3D human-skeleton sequences for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3054–3062.
[43] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multitask learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.
[44] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4597–4605.
[45] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," 2015, arXiv: 1511.04119.
[46] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.
[47] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj, "Beyond Gaussian pyramid: Multiskip feature stacking for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 204–212.
[48] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural SVM," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1422–1431, Aug. 2013.
[49] J. Wu and D. Hu, "Learning effective event models to recognize a large number of human actions," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 147–158, Jan. 2014.
[50] S. Narayan and K. R. Ramakrishnan, "A cause and effect analysis of motion trajectories for modeling actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2633–2640.
[51] S. Ramasinghe and R. Rodrigo, "Action recognition by single stream convolutional neural networks: An approach using combined motion and static information," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 101–105.
[52] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1290–1297.

Amin Ullah (S'17) received the bachelor's degree in computer science from the Islamia College Peshawar, Peshawar, Pakistan. He is currently working toward the M.S. degree leading to the Ph.D. degree in digital contents with the Intelligent Media Laboratory, Sejong University, Seoul, South Korea.

His research interests include human actions and activity recognition, sequence learning, image and video analysis, and deep learning for multimedia understanding.

Khan Muhammad (S'16–M'18) received the bachelor's degree in computer science with a focus on information security from Islamia College Peshawar, Peshawar, Pakistan, in 2014, and the M.S. leading to the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea, in 2018.

He is an Assistant Professor in the Department of Software, Sejong University, South Korea. He has authored more than 50 papers in peer-reviewed international journals and conferences, such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, FUTURE GENERATION COMPUTER SYSTEMS, *Neurocomputing*, the IEEE ACCESS, the *Journal of Medical Systems, Biomedical Signal Processing and Control, Multimedia Tools and Applications*, *SpringerPlus*, *KSII Transactions on Internet and Information Systems*, MITA 2015, PlatCon 2016, FIT 2016, ICNGC 2017, and ICNGC 2018. He is an active reviewer of more than 30 reputed journals and is involved in the editing of several special issues. His research interests include information security, image steganography, video summarization, computer vision, and video surveillance.

Javier Del Ser (M'07–SM'12) received the first Ph.D. degree in telecommunication engineering (*Cum Laude*) from the University of Navarra, Pamplona, Spain, in 2006, and the second Ph.D. degree in computational intelligence (*Summa Cum Laude*) from the University of Alcala, Madrid, Spain, in 2013.

He is currently a Research Professor in data analytics and optimization with TECNALIA (Spain), a Visiting Fellow with the Basque Center for Applied Mathematics (Spain), and an Adjunct Professor with the University of the Basque Country UPV/EHU. He has published more than 220 articles and conference contributions, cosupervised more than ten Ph.D. theses, edited four books, and coinvented six patents. His research interests include the use of descriptive, prescriptive, and predictive data mining and optimization in a diverse range of application and sectors, such as energy, transport, telecommunications, industry and security, among others.

Sung Wook Baik (M'16) 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, IL, USA, in 1992, and the Ph.D. degree in information technology engineering from George Mason University, Fairfax, VA, USA, in 1999.

He was with a Senior Scientist of the Intelligent Systems Group, Datamat Systems Research Inc. from 1997 to 2002. In 2002, he joined the faculty of the College of Electronics and Information Engineering, Sejong University, Seoul, South 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 at Sejong University. His research interests include computer vision, multimedia, pattern recognition, machine learning, data mining, virtual reality, and computer games.

Victor Hugo C. de Albuquerque (M'17) received the graduation degree in mechatronics technology from the Federal Center of Technological Education of Ceará, Fortaleza, Brazil, in 2006, the M.Sc. degree in tele-informatics engineering from the Federal University of Ceará, Fortaleza, Brazil, in 2007, and the Ph.D. degree in mechanical engineering with emphasis on materials from the Federal University of Paraíba, João Pessoa, Brazil, in 2010.

He is currently an Assistant VI Professor with the Graduate Program in Applied Informatics at the University of Fortaleza, Fortaleza, Brazil. He has experience in computer systems, mainly in the research fields of applied computing, intelligent systems, visualization and interaction, with specific interest in pattern recognition, artificial intelligence, image processing and analysis, Internet of Things, Internet of Health Things, as well as automation with respect to biological signal/image processing, image segmentation, biomedical circuits, and human/brain–machine interaction, including augmented and virtual reality simulation modeling for animals and humans.