This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

Cloud-Assisted Multi-View Video Summarization using CNN and Bi-Directional LSTM

Tanveer Hussain, *Student Member, IEEE*, Khan Muhammad, *Member, IEEE*, Amin Ullah, *Student Member, IEEE*, Zehong Cao, *Member, IEEE*, Sung Wook Baik, *Senior Member, IEEE*, Victor Hugo C. de Albuquerque, *Senior Member, IEEE*

***Abstract*—The massive amount of video data produced by surveillance networks in industries instigate various challenges in exploring these videos for many applications such as video summarization (VS), analysis, indexing, and retrieval. The task of multi-view video summarization (MVS) is very challenging due to the gigantic size of data, redundancy, overlapping in views, light variations, and inter-view correlations. To address these challenges, various low-level features and clustering based soft computing techniques are proposed that cannot fully exploit MVS. In this article, we achieve MVS by integrating deep neural network based soft computing techniques in a two tier framework. The first online tier performs target appearance based shots segmentation and stores them in a lookup table that is transmitted to cloud for further processing. The second tier extracts deep features from each frame of a sequence in the lookup table and pass them to deep bi-directional long short-term memory (DB-LSTM) to acquire probabilities of informativeness and generate a summary. Experimental[1] evaluation on benchmark dataset and industrial surveillance data from YouTube confirms the better performance of our system compared to state-of-the-art MVS.**

***Index Terms*—Artificial Intelligence, Cloud Computing, Convolutional Neural Networks, Industrial Surveillance, Multi-view Videos, Soft Computing, Video Summarization.**

I. Introduction

THE tremendous amount of video data generated by camera networks installed at industries, offices, and public places meet the requirements of Big Data. For instance, a simple multi-view network with two cameras, acquiring video from two different views with 25 frames per second (fps), generates 180,000 frames (90,000 for each camera) for an hour. The surveillance networks acquire video data for 24 hours from multi-view cameras, thereby making it challenging to extract useful information from this Big Data. It requires significant effort when searching for salient information in such huge-sized 60×60 video data. Thus, automatic techniques are required to extract the prominent information present in videos without involving any human efforts. In the video analytics literature, there exists several key information extraction techniques such as video abstraction [1], video skimming [2], and VS [3]. VS techniques investigate input video for salient information and create a summary in the form of keyframes or short video clips that represents lengthy videos. The extracted keyframes assist in many applications such as action and activity recognition [4, 5], anomaly detection [6], and video retrieval [7].

The domain of VS is broadly divided into single-view VS (SVS) and MVS. SVS summarizes a single view video and generates a short and representative output. SVS is a hot area of research and there are several traditional [8] and learned [9] features based techniques for smart surveillance in industries and various other scenarios. For instance, Mahmoud et al. [10] proposed an SVS pipeline by using density-assisted unsupervised machine learning technique known as spatial clustering to generate a video summary. Similarly, an SVS approach is presented in [11] using clustering along with semantical, emotional, and shoot-quality clues for user-generated summaries. Fei et al. [12] used the fused score of memorability and entropy to generate final summary. Majority of the research in literature focus on SVS due to its simplicity as compared to MVS.

MVS acquires the representative frames of multi-view videos (MVV) by considering both inter-view and intra-view correlations. The fundamental workflow of MVS follows three steps including pre-processing, features extraction, and post-processing for summary generation. The pre-processing of MVV involves redundancy removal techniques for video segmentation. The features extraction aims at object detection or tracking, and post-processing involves different learning or matching based techniques for computing inter-view correlations and final summary generation. The final generated summary is useful in many applications [13-15] including indoor and outdoor CCTV automatic monitoring [16] for activities and events detection [17]. Furthermore, it can be utilized for post-accident scenarios investigation, retrieval applications, and the salient information can assist in many diverse domains such as law enforcements, sports, and entertainment. Similarly, MVS can be used in the field of virtual reality for creating single 360° image view captured through various cameras.

Manuscript received December 25, 2018; Accepted: XXX, Published: XXXX. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2019R1A2B5B01070067). This paper was recommended by Associate Editor XYZ. (*Corresponding author: Sung Wook Baik*)

Tanveer Hussain, Amin Ullah, and Sung Wook Baik are with Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, Republic of Korea (Email: tanveer445@ieee.org, aminullah@ieee.org, and sbaik@sejong.ac.kr)

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

Zehong Cao is with the Discipline of ICT, University of Tasmania, TAS 7001, Australia (Email: zehong.cao@utas.edu.au).

Victor Hugo C. de Albuquerque is with Graduate Program in Applied Informatics at the Universidade de Fortaleza, Fortaleza/CE, Brazil (Email: victor.albuquerque@unifor.br)

[1] https://github.com/tanveer-hussain/MVS-using-CNN-and-LSTM

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

MVS literature follows different machine learning, deep learning, saliency, and motion based techniques. For instance, Fu et al. [18] used random walks clustering applied on spatio-temporal shot graphs. A hypergraph-based representation is used for computing correlations among different views which is then converted into spatio-temporal graph. A concept of text summarization, known as maximal marginal relevance (MMR) is used for MVS by [19]. The main idea of video MMR is to reward relevant keyframes and penalize the redundant ones. Inspired from the aforementioned work of using video-MMR for MVS, Ou et al. [20] proposed an online distributed MVS by integrating MMR with bandwidth-efficient distributed algorithm for finding K-nearest and farthest neighbors. They proposed an online keyframes generation method by exchanging small data between multiple sensors and the server. They utilized K-means clustering for final keyframes selection. Mahapatra et al. [21] proposed a scheme to produce a video synopsis for MVV based on human actions in both indoor and outdoor scenarios. Human actions are classified using support vector machine to create a video synopsis. Sparse coding is used by Panda et al. [22] for SVS and MVS, where they computed inter- and intra-view correlations in joint embedding space. The authors used "BVLC CaffeNet [23]" pre-trained CNN model to extract features for computing intra-view correlations and preserved them in embedding space. Finally, they used latent subspace clustering to utilize the features in embedding space for summary generation. Panda et al. [24] computed two proximity matrices and accumulated inter- and intra-view correlations to calculate pairwise Euclidean distances between frames. They finally used sparse representative selection over learned embedding for summary generation. In another MVS research [25], inter- and intra-view correlations are computed via sparse coefficients. It contains a coefficient matrix, which represents relative importance of segmented shots. The shots with highest importance score are considered as a part of the final summary by using clustering technologies.

It can be observed from the MVS literature that majority of the methods are based on low-level features and apply traditional machine learning techniques for summary generation. In addition, these methods generate summary locally without involving cloud computing [26] for faster MVS and instant detailed analysis. Therefore, to target these challenges, this article introduces an efficient framework for MVS that performs light weight computation locally and costly processing on cloud. Our key contributions are summarized as follows:

1. Shots segmentation is considered as a critical step for MVS methods. The currently employed methods are not well-suited for industrial settings, thus, in this paper, we present a target appearance-based shots segmentation mechanism. In industrial surveillance, different targets including human and vehicles are very important for further analysis. Thus, our framework performs target appearance based shots segmentation in a cost-effective manner.
2. The stiffest and important pre-processing task in MVS is computing inter-view correlations, which is not well-addressed yet with better results for MVS. In this framework, we introduce a novel and efficient concept of lookup table for computing inter-view correlations. Lookup table is acquired by storing the segmented shots with targets in a timely manner and synchronized for all videos. The key contribution of our proposed framework is computing inter-view correlations without involving any extra processing unlike other traditional MVS methods.
3. The mainstream MVS techniques are based on low-level features that use clustering or template matching algorithms to generate summary. These traditional methods are not effective to generate a representative summary for industrial surveillance videos. To overcome this issue, we employ sequential learning for summary generation using deep features and DB-LSTM, which outputs the probabilities of informativeness for a sequence of frames. Thus, decision based on the information present inside video sequences using learned features makes our system a suitable candidate in terms of efficiency and accuracy for MVS especially in industrial environments.
4. Industrial surveillance generates huge amount of data daily, needing a considerable amount of computational complexity and time for exploration. To efficiently analyze only important data, we transmit the segmented shots of video to cloud for MVS. Cloud computing assists in efficient and precise processing of video data and the generated summary can be processed for advanced analysis such as abnormal activities and violence recognition without any transmission delay.

The rest of the paper is structured as follows. Section II describes all the major components of our framework in detail. Section III describes the experimental results and discussion. In Section IV, this work is concluded with recommendations for further research.

II. Proposed MVS Framework

In this section, the overall working of the proposed framework is discussed in detail. The proposed approach is divided into two tiers. First tier is online which performs shots segmentation and preserves the shots in a lookup table in a timely manner. Second cloud-based tier extracts learned features from these shots and analyzes sequential patterns through DB-LSTM model for final summary generation. The overall framework with both tiers is visualized in Fig. 1 and explained in Section A and B.

A. Online Tier

This tier captures multi-view videos and passes them to the object detection trained CNN model, which detects the desired targets and segments the video into shots. A novel concept of lookup table is introduced in this framework to store the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

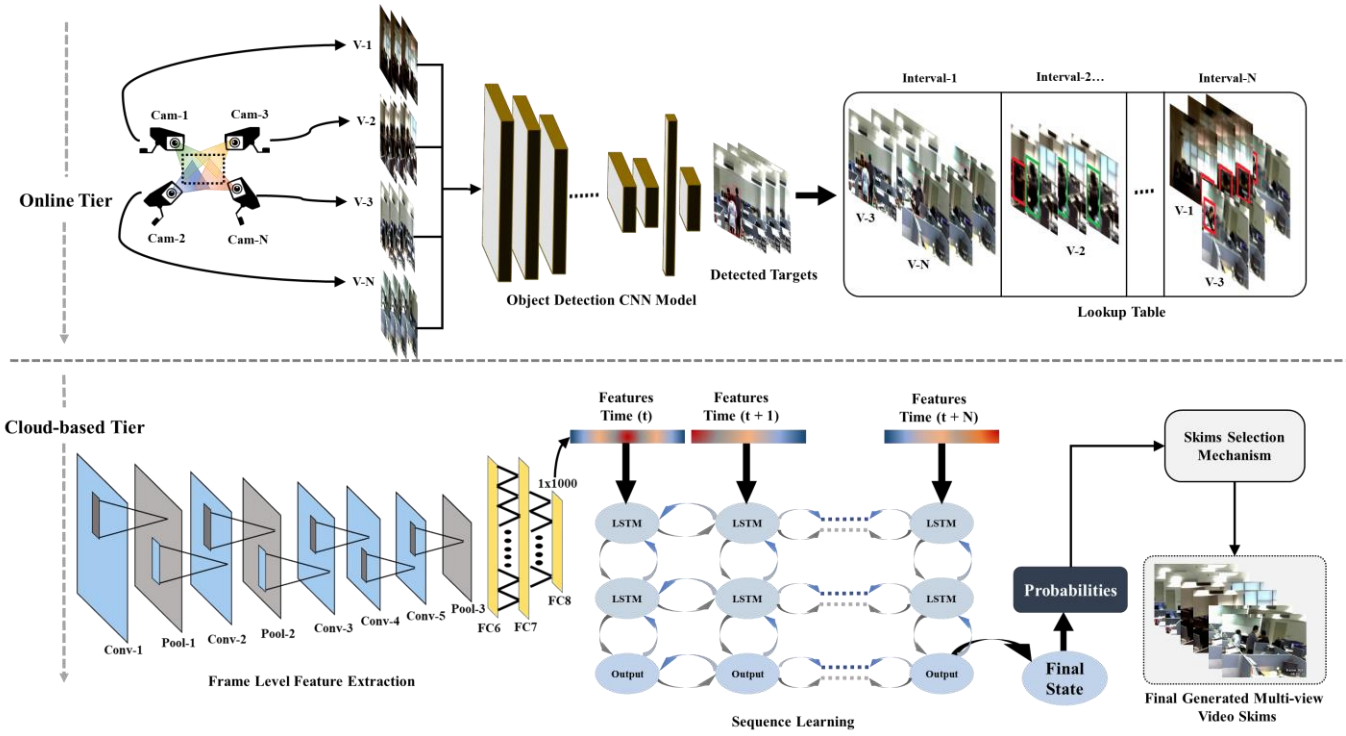


Figure 1: The proposed two tier cloud-assisted framework for MVS. Online tier: it segments target appearance-based shots and stores them in a lookup table. Cloud-based tier: data in the lookup table is transmitted to cloud where deep features are extracted from each video frame and passed to DB-LSTM for sequence learning. DB-LSTM outputs the probabilities of information present in these sequences. The sequences with highest probabilities of informativeness are integrated together to generate final summary.

segmented shots synchronously and transmit them to next tier for further processing. The online tier performs all computations in real-time with 15 fps.

1) *Shots Segmentation*

Shots segmentation is an important step for VS and a summary is heavily dependent on it. A good shots segmentation technique ensures better quality of summary. It is a difficult step because the contents of video may change abruptly without any sequence and it is necessary to detect these changes for separation of one shot from the other. In VS literature there are various approaches for shots segmentation. For instance, [18] performs activity based shot segmentation while an another approach presented in [27] detects shots boundaries based on the motion in video data. Panda et al. [24, 25] utilizes Spatio-temporal C3D features for video representation in shots. Similarly, Muhammad et al. [9] exploits deep features for shots segmentation. The discussed techniques are not suitable for surveillance camera networks particularly in industrial scenarios because they pay no attention to human or vehicles while generating a summary. Further, these cameras generate hours long videos with different events. In these videos, human perform different actions and activities and vehicles may encounter some uncertain situations which confirms that the salient objects for industrial surveillance are human and vehicles. Thus, the video summary must include only those shots that contain human and vehicles. Therefore, considering the challenges of surveillance videos in industrial environments, we exploit CNN for object detection to segment video into shots.

a) Fine-tuning Yolo Object Detection Model

Recently, many studies investigated CNNs for various applications such as security [28, 29], medical [30], and action recognition [4, 5]. In this article, motivated from the aforementioned studies we use CNN for target detection. Normally object detection models are trained on objects from general categories with no focus on surveillance where objects are at larger distances and blurry, making their detection comparatively challenging. Therefore, we fine-tuned an existing object detection CNN architecture for our desired classes i.e., human and vehicles in surveillance scenarios. The fine-tuning process assists in increasing the accuracy of detection in surveillance scenarios even if targets are far away from camera. We utilize YOLO [31] object detection model that is extremely fast and accurate for targets detection. For annotation, we collected data from YouTube and MVS datasets including Office [18] and Bl-7F [32]. We manually labelled target's locations in each frame. We trained our model for 1000 epochs and saved the weights when the average loss reached 0.008. YOLO treats the object detection problem as regression and simultaneously predicts region of interests for multiple classes, which is most precise and efficient. YOLOv3 contains a sequence of 53 convolutional layers and uses Leaky Rectifier after each convolutional layer as a non-linear activation function. The size of max-pooling layer is 2 × 2 with stride 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

The major benefit of YOLO over the other object detectors is its fast processing with better accuracy.

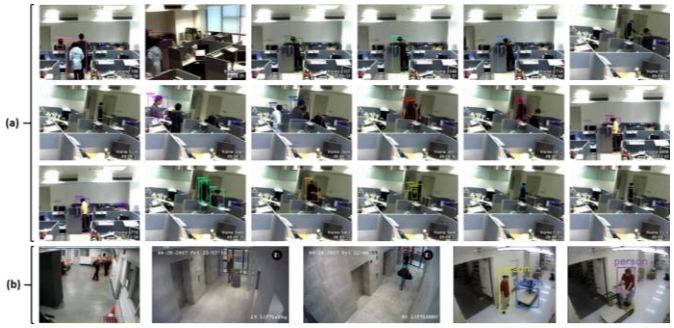


Figure 2: Sample results of our shots segmentation mechanism. (a) Office dataset and (b) industrial surveillance. These shots are stored in lookup table and are transmitted to cloud for further processing. Sample results show that the proposed system segmented only the salient shots that can be used in further steps for MVS.

b) Shots Selection Mechanism

The proposed mechanism uses the concept of multi-threading to perform shots segmentation parallelly. To store the synchronized shots of targets for all views, it is mandatory to process videos in parallel. Videos from multi-view camera network are acquired and given as input to our system. For every input video, a separate thread is created which processes it via multi-threading for the shots selection mechanism in order to execute this process concurrently as per requirement of the system for the lookup table. The frames with targets are stored in the lookup table for all the views at time interval ***T***. The time interval ***T*** contains ***n*** number of frames which is set manually. In the proposed framework, we set the value of n to 15 because we are computing the videos with 15 fps and it ensures that no important information is lost. The effectiveness of our shots segmentation approach can be observed from Fig. 2 where our system selected the shots from surveillance and Office dataset videos with targets only.

B. Cloud-based Tier

The cloud-based tier is responsible for final summary generation. The segmented shots in the lookup table are received on the cloud server. Visual frame-level deep features from these shots for a sequence of 15 frames are extracted and are inputted to DB-LSTM. Finally, summary is generated on the basis of probabilities predicted from DB-LSTM. Each step is explained in detail in subsequent sections.

1) Deep Features Extraction

This section describes the importance and technical details about the deep features extraction mechanism of our proposed framework. It is very difficult task to classify the nature of an image on the basis of information present in it. Therefore, we investigated CNN and LSTM to decide whether a sequence of frames is informative or not. There are some low-level features based concepts for finding the informativeness of frames such as entropy [9] and complexity [33]. Entropy designates the informativeness of a frame, higher is the value of entropy more informative is the frame and vice versa. Similarly, higher complexity value for a frame indicates the more informativeness of the frame and vice versa. The discussed methods for information computation are based on low-level features and not suitable for MVV having camera instability and illumination. To overcome these challenges, we employed CNN and exploited a DB-LSTM model to compute informativeness of a sequence of frames in every sort of environment. We conducted experiments using light-weight CNNs such as SqueezeNet and MobileNetV2 [34], but the obtained results were not convincing. The dense layers of these CNNs fail to represent the discriminative features in the sequence of frames to fed them into DB-LSTM model. Therefore, we utilized the learned features of AlexNet [35] CNN model. It is trained on a large-scale ImageNet [36] dataset for classification. However, it is proven from recent studies [37, 38] that the initial layers of pre-trained CNN model process an image using convolutional kernels, which can extract local semantic representation of an image. While the parameters in deeper layers of pre-trained CNN model can capture global and discriminative patterns in an image which can be used for many other tasks. We have utilized the features of a deeper layer for frame representation in our sequence learning process to compute the informativeness. The final layer of AlexNet architecture outputs 1×1000 feature vector for a single frame. We have extracted features from the sequence of 15 frames of video and the obtained feature vector is input to DB-LSTM for further processing which is explained in the next section. Although deep features extraction requires a lot of computational power but in our proposed framework, we perform this step on cloud to save time and resources of local servers.

2) Learning via DB-LSTM

Recurrent neural networks (RNN) take input at multiple time steps and analyze the sequential patterns from the previous data unlike traditional deep neural networks which consider only a single input. Traditional RNN has two sources of input which takes a new input at each time step and receives the output of previous time step. As simple RNN takes input and generates output at each time step, therefore it falls into vanishing gradient problem which forgets the effect of a longer sequence. Thus, RNN is not a good option for finding temporal-sequential patterns in long-term time series data. LSTM has the solution to vanishing gradient problem of simple RNN by exploiting different gates. It has analogue gates (input, forget, and output gates) that help in learning long-term sequential information. The LSTM takes decision using element-wise multiplication and sigmoid activations for gates to be opened or closed. The "Forget gate" in LSTM chooses information to throw away from memory cell state and outputs 0 or 1 for each number in the cell state where 1 means "keep this" and 0 represents "get rid of this". The "input gate" decides where the input needs to be taken on this time step or not while a *tanh* activation layer creates a vector of new candidate values (C_t) that is added to the cell memory. Next, the old cell state at (C_{t-1}) is updated to new cell state (C_t) by multiplying old state with new input followed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

by an activation function. In the proposed framework, we stacked two layers of LSTM on top of each other to form DB-LSTM that helps in learning long-term changes. In deep or multi-layer LSTM the output of the bottom layer is inputted to the top layer and previous state of the same layer. The bi-directional structure of LSTM processes the sequence in forward and backward directions which help to learn from the past and future changes in the sequence. For more explanation about DB-LSTM, readers can refer to [39]. Usually, LSTM gives output at different time steps which is decided by sigmoid activation of the output gate. However, we used the output of the final state of LSTM that exhibits full processed sequence which is then fed to Softmax classifier for final prediction.

Deep features are extracted from a sequence of 15 frames at time *T* and forward propagated to DB-LSTM for learning whether a sequence of frames is informative or not. The employed DB-LSTM contains 256 memory cells and is trained for 500 epochs, having 60% training, 20% validation, and 20% testing samples. The optimal learning rate after several experiments is selected as 0.001 for cost optimization by using stochastic optimization [40]. Finally, we obtain a trained model that is able to classify sequences as informative or non-informative which helps in final summary generation.

3) Summary Generation Mechanism

Although the final nominated skims are important enough, yet there is a possibility that some visually similar sequences are selected as summarized skims. This problem makes the final summary generation step challenging which is tackled in the proposed technique by selecting only those frames as final summary whose probability is maximum for the informativeness class. Thus, the best fit frames with highest probabilities are considered in the post-processing step which ensures a diverse and representative summary. In this step, the frames with maximum probabilities are combined together to generate a single summarized video representative of all the views. The final summaries with probability scores can be observed from Fig. 3 where the frames with maximum probabilities from a sequence are part of the final summary.

III. Experimental Results and Discussion

In this section, our proposed MVS framework is experimentally evaluated using two perspectives i.e., objective and subjective. The results are generated using different evaluation metrics including precision, recall, and F1-measure [32]. The proposed approach is evaluated using Office dataset [18] that is publicly available to research community. Furthermore, the experimental results are compared with other state-of-the-art MVS methods to prove the effectiveness of our proposed framework.

A. Experimental Settings

The proposed system is implemented in Python language version 2.7 using Spyder integrated development environment. The very famous deep learning framework Caffe is utilized for extracting learned features from AlexNet [35] model. Another deep learning framework TensorFlow is used for the implementation of DB-LSTM. Furthermore, we used two computers with different configurations for assessment of our proposed system. First configuration is used for online tier and the second one is for the cloud-based tier. The online tier is executed in real-time with 15 fps to make the process efficient and effective. Description of the systems is given in Table I.

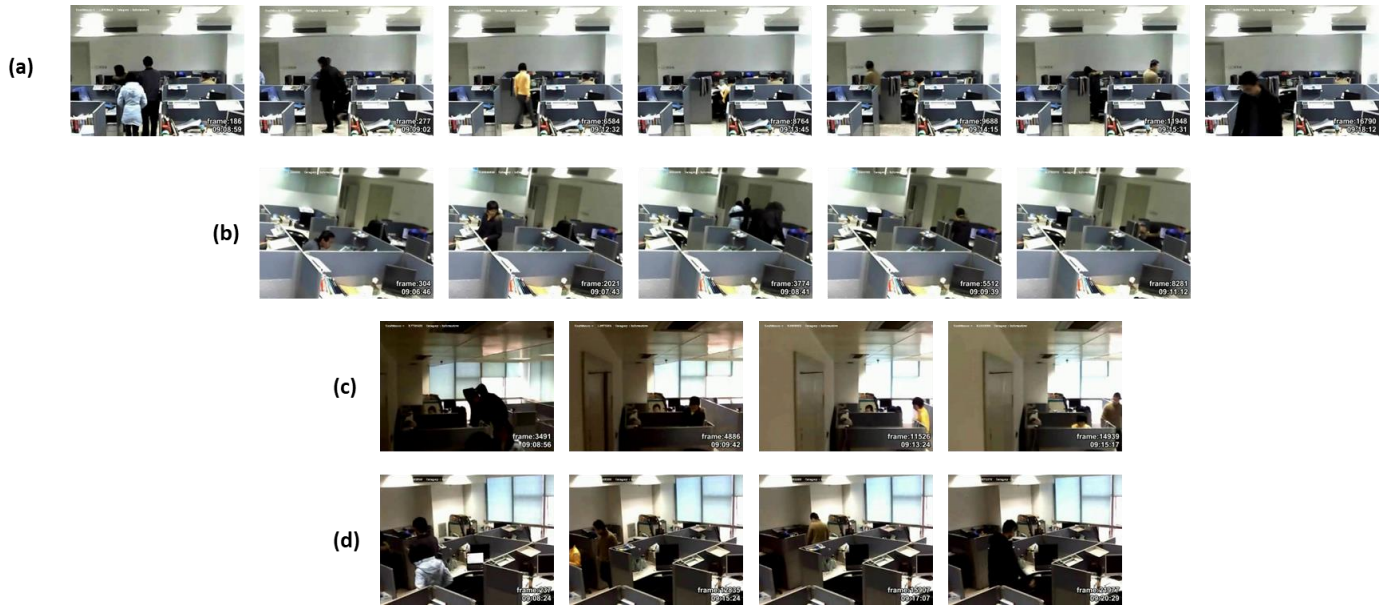


Figure 3: Sample frames for each view generated by our proposed framework. (a) A, B, C walk out, B put on coat, D walk in, D sit down, E walk out, E walk in, B walk away, (b) B sit down at middle sit, A look at the middle sit, A, B, C walk out, A walk in, A sit down, (c) B put on coat, A stand up, D walk in, E walk out, and(d) A, B, C walk out, E walk out, E walk in, B walk away.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

Figure 4: Sample keyframes from generated skims using our proposed framework for industrial surveillance videos. First row has the keyframes of a video in which an employee tries to fix some leakage problem in the pipes where he creates mess by leaking the whole pipe. The same idea is revealed in the keyframes and can be observed very easily from the start to end. He brings a small ladder to fix the leakage which is not reachable and then he brings another big ladder to reach the pipe. In the last frames the employee creates a big hole in the pipe where water sprints and he falls. In the second row, last two columns represent keyframes of a video where an employee falls from a certain height and other workers comes for his rescue. Third row contains the keyframes where certain stuff falls from the roof on workers and they try to escape. The last row represents an industry indoor video where an employee is doing some regular cleaning activity. As visualized in figure the frames with only employees doing certain activity or going out of room are extracted by our proposed framework.

B. Datasets and Data Acquisition

The proposed system is evaluated using publicly available benchmark Office [18] dataset for MVS and different single-view industrial surveillance videos from YouTube. The target appearance is achieved through YOLOv3 object detection model. We fine-tuned the existing model for only two classes by providing extra data from surveillance videos containing our targets. Furthermore, we also collected different sequential short clips from Office and Action recognition [41, 42] datasets for walking and sitting class. The data also contains industrial surveillance videos downloaded from YouTube. We acquired 400 short videos for informative class and 300 for non-informative class. The sequences are marked as informative and non- informative based on the presence of an event. This data is fed into AlexNet CNN model for features extraction which are forwarded to DB-LSTM for sequence learning. The MVS dataset used for evaluation is discussed below. The details of surveillance videos downloaded from YouTube for training purposes are given in Table II.

TABLE I
PERFORMANCE MATRIX FOR OFFICE MVS DATASET COMPUTED FOR EACH VIEW INDIVIDUALLY

View	Precision	Recall	F1 Score	Event Recall
office-0	0.91	0.88	0.89	0.88
office-1	0.94	0.85	0.89	0.85
office-2	0.94	0.88	0.91	0.88
office-3	0.92	0.85	0.88	0.87

1) Office Dataset [18]

It is one the most popular datasets in MVS literature. Office is the first dataset in MVS research and is used by most of the MVS techniques for experimental evaluation. This dataset is created using 4 stably-held cameras in an office with no synchronization among them. There are various light conditions at different views in this dataset. Thus, these challenges in Office dataset makes it difficult for evaluation of MVS methods, leading to their poor achievement in terms of accuracy to date.

C. Objective Evaluation

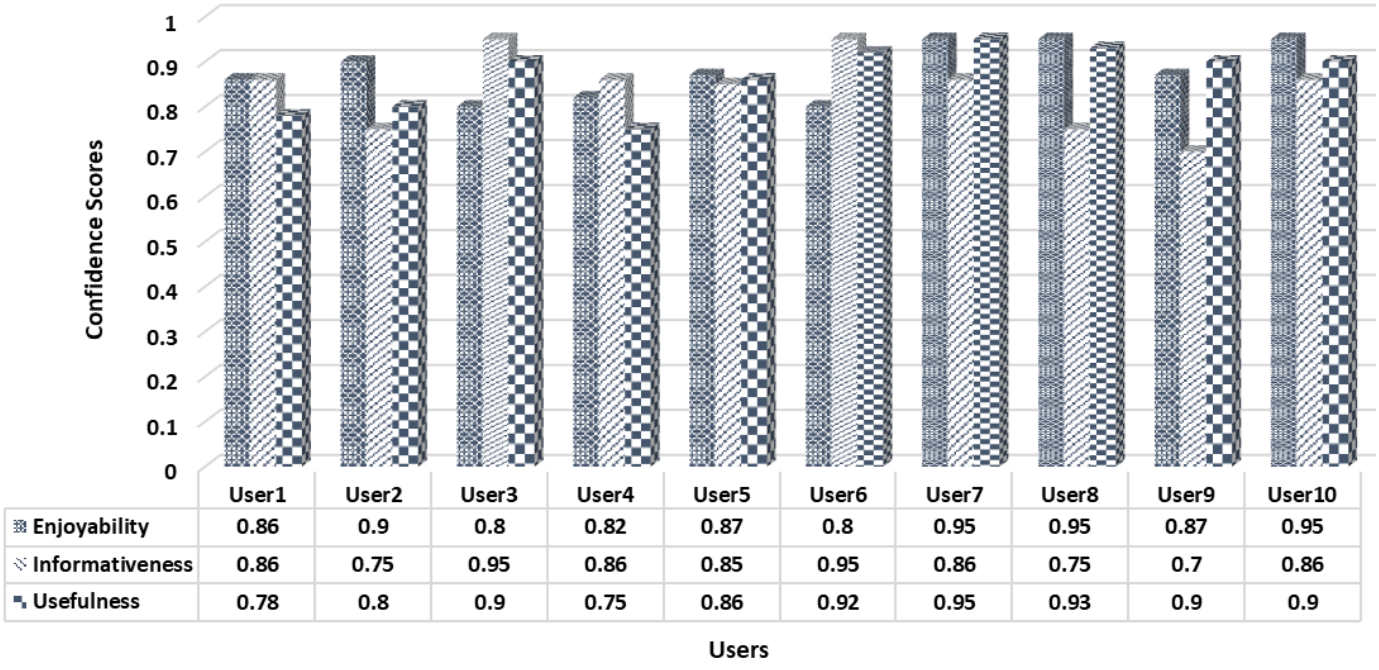
This section explores the comparison of our proposed framework with existing MVS research methods for the Office dataset to show its better performance. The metrics used for comparison are precision, recall, and F1-score for a fair comparison of our proposed framework with state-of-the art. Precision directs the aptitude of a VS technique to remove useless frames. Recall indicates the capability of a VS technique to keep the salient information in the generated summary. Therefore, to prove the better performance of our framework we used these metrics for evaluation. Sample keyframes from different skims of all views of the Office dataset are represented in Fig. 3. In Fig. 4 the representative frames for industrial surveillance videos are visualized. The results in terms of events analysis from Office dataset in each view are given in Table III. It is clear from the last column of Table III that recall value ranges between 0.85 and 0.88 that indicates the better event detection of our proposed framework.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

Similarly, it can be observed from Table IV that we compared our framework with seven existing approaches. The most recent candidate paper for comparison is published in 2017 and we compared our results with their method [25]. We also compared our framework with other approaches based on MVS. A gap in the value of precision can be observed in Table IV where our system is lagging behind from the state-of-the-art [25] due to the presence of some redundant frames in the final summary. The proposed approach outruns all the existing methods and F1-score is improved 0.1 compared to [25], which is considered as the best among all MVS approaches. The last column indicates the event recall value which is increased by 1.2 from [43]. Thus, the objective evaluation reveals that our system outruns the state-of-the-art and proves to be a better option for MVS in industries.

TABLE II
DETAILED DESCRIPTION OF SURVEILLANCE VIDEOS WITH DIVERSE TYPES OF ACTIVITIES

S.no	Title	Link	Description
1	Target Warehouse Forklift Follies	https://www.youtube.com/watch?v=2tma-zFkDEY	Industry video with a person fixing leakage of water pipes. The video resolution is very low and contains noise.
2	Couple stealing fuel bungle a getaway from Mt Warren Park petrol station	https://www.youtube.com/watch?v=NTMT0KDp7SM	Low resolution video with persons and vehicles at a petrol pump. A person attempts to steal car from a lady.
3	Anchorage Jail surveillance video of Larry Kobuk death	https://www.youtube.com/watch?v=a1Zo3zfvZwY&t=928s	CCTV footage of a prisoner's death from different views with variable light conditions and low resolution.
4	EXCLUSIVE: Chilling CCTV footage of Karabo Mokoena's killer moments before he disposes of her body	https://www.youtube.com/watch?v=pv-4WQBy0UI	Surveillance of disposal of a human body. Video contains multiple subjects from different views.
5	Industrial Warehouse Thief is Arrested in Los Angeles, California	https://www.youtube.com/watch?v=fZzbef-eu-Q	A thief in an industry trying to steal some materials. The person is very far from the camera and video is captured at day time with extreme lighting conditions
6	Man Attempts to Steal Supplies and is Arrested - Industrial Security	https://www.youtube.com/watch?v=hdtquo_n2BE	A man is arrested in this video who is trying to steal supplies from an industry.
7	Sometimes Security Cameras catch a gem!	https://www.youtube.com/watch?v=4i_GFrlaStQ	A surveillance video with person and vehicle. The camera is very far from the targets and has low resolution
8	Surveillance video of van attack on Dutch newspaper office	https://www.youtube.com/watch?v=zUmu9v3J3KA	A person in a vehicle hits the office building and then fires it. The video is made in dark and the person and vehicle is far from the camera.
9	Unfukinbelievably tragic workplace accidents. Mature content	https://www.youtube.com/watch?v=j776Vfy1hzM&t=64s	A video with various surveillance accidents. Some clips from this whole video are used for experiments.
10	Target Employee driving wave...badly	https://www.youtube.com/watch?v=6blZ12ZSJtM	Employees working in a lobby, the main focus is on a worker who is cleaning. Other workers pass by him. The resolution of the camera is low.



	User1	User2	User3	User4	User5	User6	User7	User8	User9	User10
Enjoyability	0.86	0.9	0.8	0.82	0.87	0.8	0.95	0.95	0.87	0.95
Informativeness	0.86	0.75	0.95	0.86	0.85	0.95	0.86	0.75	0.7	0.86
Usefulness	0.78	0.8	0.9	0.75	0.86	0.92	0.95	0.93	0.9	0.9

Figure 5: Statistical data of 10 users for the values of enjoyability, informativeness, and usefulness. Each user is asked to assign score between 0 and 1. All the participants assigned good scores to the usefulness of the generated summaries. An average enjoyability for all the participants is close to 0.9. The informativeness score assigned by many participant's ranges between 0.7 and 0.95

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

D. Subjective Evaluation

Due to the subjective nature of VS domain, only objective evaluation of a method is not enough to show its validity. To further evaluate the better performance of our proposed framework, we carried out a user study. In this study, 10 participants are provided with the Office dataset videos, ground truth for the office dataset, and the summary generated by our proposed framework. Furthermore, we provided them the summary produced from industrial surveillance without any ground truth to decide whether the generated summary is rich of information or not. The candidates who participated in the survey are Master and PhD students familiar to the field of computer vision and have experience of evaluation of different methods. The age of participants ranged from 22 to 28 years. In order to carry out an unbiased survey, we asked the participants to submit the results after two days so that they can comfortably answer to the questions without any pressure of time. The questions for evaluation of our framework are almost the same as provided by [18]. The users were asked to assign score between 0 and 1 to all questions. The overall statistics of users are given in Table V, while individual scores are visualized in Fig. 5.

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED FRAMEWORK WITH STATE-OF-THE-ART ON OFFICE DATASET.

Methods	Precision	Recall	F1 Score	Event recall
[18]	1	0.61	N/A	N/A
[43]	0.41	0.63	0.50	0.75
[27]	1	69.2	81.79	N/A
[32]	27.5	75.6	0.40	0.45
[22]	1	0.73	81.48	N/A
[24]	1	0.70	0.81	N/A
[25]	**1**	0.81	0.89	N/A
Ours	0.93	**0.860.**	**0.90**	**0.87**

E. Time Complexity Analysis

The details about time complexity for summary generation on local computer and cloud server are given in Table VI. It can be observed from Table VI that it is feasible to generate summary on cloud server in terms of processing time. The specification of both local and cloud servers are the same but the only difference is that cloud server employs GPU computing. The video used for testing is Office-1 having 235.8 Mb size, 9 minutes 4 seconds length, and 15 fps.

TABLE IV
TIME ANALYSIS AND COMPARISON OF THE PROPOSED FRAMEWORK FOR SUMMARY GENERATION ON CLOUD SERVER AND LOCAL COMPUTER

Platform	Processing time (seconds)
Local server/ Personal computer	2048.88
Cloud server/GPU	343.01

IV. CONCLUSIONS AND FUTURE RESEARCH

In today's modern era surveillance networks are installed almost everywhere. These networks generate 24 hour videos on daily basis with significant redundancy, leading to wastage of storage resources as well as making their analysis difficult. Motivated by these challenges, we proposed an effective CNN and DB-LSTM based MVS framework. Our framework first segments the multi-view videos into shots based on the appearance of human and vehicles. The segmented shots acquired in online tier are stored in a lookup table with timestamp and then transmitted to cloud for further processing. Thus, our system only transmits the useful and desired data to cloud and plays a vital role in saving bandwidth and computational resources. In the cloud-based tier, we used a CNN architecture to extract deep features from a sequence of 15 frames. We input these features to DB-LSTM which is trained to learn informative and non-informative sequence of frames and outputs probabilities for these two classes. At last, the sequences with highest probabilities of informativeness are included in the final summary. Extensive experiments and comparisons with other state-of-the-art techniques verify the dominance of our system.

Currently, we used heavy weight CNN model that we want to replace by optimized deep learning model with similar or higher accuracy. Furthermore, we want to extend our framework through some supplementary analysis steps such as activity recognition and analysis for instant reporting about the abnormal activities.

References

[1] S. Zhang, X. Li, S. Hu, and R. R. Martin, "Online Video Stream Abstraction and Stylization," *IEEE Transactions on Multimedia*, vol. 13, pp. 1286-1294, 2011.

[2] L. Zhang, L. Sun, W. Wang, and Y. Tian, "KaaS: A Standard Framework Proposal on Video Skimming," *IEEE Internet Computing*, vol. 20, pp. 54-59, 2016.

[3] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*, 2016, pp. 766-782.

[4] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. Van Gool, "stagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[5] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, and Y.-D. Zhang, "Multi-domain and multi-task learning for human action recognition," *IEEE Transactions on Image Processing*, vol. 28, pp. 853-867, 2019.

[6] T. Wang, M. Qiao, Z. Lin, C. Li, H. Snoussi, Z. Liu, *et al.*, "Generative Neural Networks for Anomaly Detection in Crowded Scenes," *IEEE Transactions on Information Forensics and Security*, pp. 1-1, 2018.

[7] H. Zhang, X. Cao, J. K. Ho, and T. W. Chow, "Object-level video advertising: an optimization framework," *IEEE Transactions on Industrial Informatics*, vol. 13, pp. 520-531, 2017.

[8] H. Liu, Y. Liu, Y. Yu, and F. Sun, "Diversified Key-Frame Selection Using Structured<inline-formula><tex-math notation="TeX">${L_{2,1}}$</tex-math></inline-formula>Optimization," *IEEE Transactions on Industrial Informatics*, vol. 10, pp. 1736-1745, 2014.

[9] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient CNN based summarization of surveillance videos for resource-constrained devices," *Pattern Recognition Letters*, 2018.

[10] K. M. Mahmoud, M. A. Ismail, and N. M. Ghanem, "Vscan: an enhanced video summarization using density-based spatial clustering," in *International conference on image analysis and processing*, 2013, pp. 733-742.

[11] B. Xu, X. Wang, and Y.-G. Jiang, "Fast Summarization of User-Generated Videos: Exploiting Semantic, Emotional, and Quality Clues," *IEEE MultiMedia*, vol. 23, pp. 23-33, 2016.

[12] M. Fei, W. Jiang, and W. Mao, "Memorable and rich video summarization," *Journal of Visual Communication and Image Representation*, vol. 42, pp. 207-217, 2017.

[13] A. S. Murugan, K. S. Devi, A. Sivaranjani, and P. Srinivasan, "A study on various methods used for video summarization and moving

object detection for video surveillance applications," *Multimedia Tools and Applications*, pp. 1-18, 2018.

[14] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description," *IET Computer Vision*, vol. 10, pp. 758-767, 2016.

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

[16] T. D'Orazio and C. Guaragnella, "A survey of automatic event detection in multi-camera third generation surveillance systems," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, p. 1555001, 2015.

[17] J. Ren, M. Xu, J. S. Smith, and S. Cheng, "Multi-view and multi-plane data fusion for effective pedestrian detection in intelligent visual surveillance," *Multidimensional Systems and Signal Processing*, vol. 27, pp. 1007-1029, 2016.

[18] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, "Multi-view video summarization," *IEEE Transactions on Multimedia*, vol. 12, pp. 717-729, 2010.

[19] Y. Li and B. Merialdo, "Multi-video summarization based on Video-MMR," in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, 2010, pp. 1-4.

[20] S. H. Ou, Y. C. Lu, J. P. Wang, S. Y. Chien, S. D. Lin, M. Y. Yeti, *et al.*, "Communication-efficient multi-view keyframe extraction in distributed video sensors," in *2014 IEEE Visual Communications and Image Processing Conference*, 2014, pp. 13-16.

[21] A. Mahapatra, P. K. Sa, and B. Majhi, "A multi-view video synopsis framework," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 1260-1264.

[22] R. Panda, A. Dasy, and A. K. Roy-Chowdhury, "Video summarization in a multi-view camera network," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, 2016, pp. 2971-2976.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675-678.

[24] R. Panda, A. Das, and A. K. Roy-Chowdhury, "Embedded sparse coding for summarizing multi-view videos," in *Image Processing (ICIP), 2016 IEEE International Conference on*, 2016, pp. 191-195.

[25] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *arXiv preprint arXiv:1706.03121*, 2017.

[26] F. Al-Turjman, M. Z. Hasan, and H. Al-Rizzo, "Task scheduling in cloud-based survivability applications using swarm optimization in IoT," *Transactions on Emerging Telecommunications Technologies*, p. e3539, 2018.

[27] S. K. Kuanar, K. B. Ranga, and A. S. Chowdhury, "Multi-view video summarization using bipartite matching constrained optimum-path forest clustering," *IEEE Transactions on Multimedia*, vol. 17, pp. 1166-1173, 2015.

[28] D. Li, L. Deng, B. B. Gupta, H. Wang, and C. Choi, "A novel CNN based security guaranteed image watermarking generation scenario for smart city applications," *Information Sciences*, 2018.

[29] T. Hong, C. Choi, and J. Shin, "CNN-based malicious user detection in social networks," *Concurrency and Computation: Practice and Experience*, vol. 30, p. e4163, 2018.

[30] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321-331, 2018.

[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.

[32] S.-H. Ou, C.-H. Lee, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "On-line multi-view video summarization for wireless video sensor network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, pp. 165-179, 2015.

[33] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool, "The interestingness of images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1633-1640.

[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211-252, 2015.

[37] J. Ahmad, K. Muhammad, S. Bakshi, and S. W. Baik, "Object-oriented convolutional features for fine-grained image retrieval in large surveillance datasets," *Future Generation Computer Systems*, vol. 81, pp. 314-330, 2018/04/01/ 2018.

[38] N. Rahim, J. Ahmad, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Privacy-preserving image retrieval for mobile devices with deep features on the cloud," *Computer Communications*, vol. 127, pp. 75-85, 2018/09/01/ 2018.

[39] A. Ullah, K. Muhammad, J. D. Ser, S. W. Baik, and V. Albuquerque, "Activity Recognition using Temporal Optical Flow Convolutional Features and Multi-Layer LSTM," *IEEE Transactions on Industrial Electronics*, pp. 1-1, 2018.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, pp. 971-981, 2013.

[42] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*, 2009, pp. 1996-2003.

[43] S.-H. Ou, Y.-C. Lu, J.-P. Wang, S.-Y. Chien, S.-D. Lin, M.-Y. Yeti, *et al.*, "Communication-efficient multi-view keyframe extraction in distributed video sensors," in *Visual Communications and Image Processing Conference, 2014 IEEE*, 2014, pp. 13-16.

Tanveer Hussain (S'19) received his Bachelor's degree in Computer Science from Islamia College Peshawar, Peshawar, Pakistan in 2017. He is currently pursuing his M.S. leading to Ph.D. degree from Sejong University, Seoul, Republic of Korea and serving as Research Assistant at Intelligent Media Laboratory (IM Lab). His major research domains are features extraction (learned and low-level features), video analytics, image processing, pattern recognition, deep learning for multimedia data understanding, single/multi-view video summarization, IoT, IIoT, and resource constrained programming.

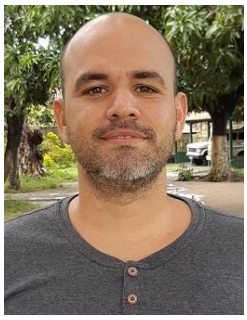
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2929228, IEEE Transactions on Industrial Informatics

Khan Muhammad (S'16–M'18) received the Ph. D degree in Digital Contents from Sejong University, South Korea. He is currently an Assistant Professor in the Department of Software, Sejong University, South Korea. His research interests include medical image analysis (brain MRI, diagnostic hysteroscopy and wireless capsule endoscopy), information security (steganography, encryption, watermarking and image hashing), video summarization, computer vision, fire/smoke scene analysis, and video surveillance. He has published over 60 papers in peer reviewed international journals and conferences in these research areas with target venues as IEEE COMMAG, Networks, TII, TIE, TSMC-Systems, IoTJ, Access, TSC, Elsevier INS, Neurocomputing, PRL, FGCS, COMCOM, COMIND, JPDC, PMC, BSPC, CAEE, Springer NCAA, MTAP, JOMS, and RTIP, etc. He is also serving as a professional reviewer for over 40 well-reputed journals and conferences.

Amin Ullah (S'17) received the bachelor's degree in computer science from the Islamia College Peshawar, Peshawar, Pakistan. He is currently pursuing the M.S. leading to Ph.D. degree with the Intelligent Media Laboratory, Sejong University, South Korea. He has published several papers in reputed peer reviewed international journals and conferences including IEEE TIE, IEEE IoTJ, IEEE Access, and Elsevier FGCS. His research interests include human actions and activity recognition, sequence learning, image and video analysis, and deep learning for multimedia understanding.

Zehong Cao (M'13) is a Research Fellow at Centre for Artificial Intelligence/School of Software, Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Australia. He conferred a dual PhD program in Information Technology from UTS, and Electrical and Control Engineering from National Chiao Tung University (NCTU). He received a master degree from The Chinese University of Hong Kong (CUHK) and a bachelor degree from Northeastern University (NEU). Currently, he is mainly focusing on the capacity of the human brain communicating and interacting with the computer and environment, at assisting and augmenting human cognition. His research interests cover signal processing, data mining, brain-computer interface, bioinformatics, fuzzy systems, neural networks, machine learning, cognitive neuroscience, optimization and clinical applications. His research objective is to exploit computational intelligence methodologies for brain-machine interfaces. Recently, with the outstanding of research performance, he was honored with the Associate Editor of IEEE Access, Guest Editor of Swarm and Evolutionary Computation, Neurocomputing, and International Journal of Distributed Sensor Networks.

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



Victor Hugo C. de Albuquerque (M'17–SM'19) 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, 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. 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.