A video summarization framework based on activity attention modeling using deep features for smart campus surveillance system

Wasim Muhammad[1], Imran Ahmed[1], Jamil Ahmad[1,2],
Muhammad Nawaz[1], Eatedal Alabdulkreem[3] and Yazeed Ghadi[4]

[1] Center of Excellence in Information Technology, Institute of Management Sciences
(IMSciences), Peshawar, Peshawar, KPK, Pakistan
[2] Department of Computer Science, Islamia College Peshawar (Chartered University), Peshawar,
Pakistan
[3] Computer Sciences Department, College of Computer and Information Sciences, Princess
Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[4] Department of Computer Science, Al Ain University, Al Ain, UAE

ABSTRACT

Like other business domains, digital monitoring has now become an integral part of almost every academic institution. These surveillance systems cover all the routine activities happening on the campus while producing a massive volume of video data. Selection and searching the desired video segment in such a vast video repository is highly time-consuming. Effective video summarization methods are thus needed for fast navigation and retrieval of video content. This paper introduces a keyframe extraction method to summarize academic activities to produce a short representation of the target video while preserving all the essential activities present in the original video. First, we perform fine-grain activity recognition using a realistic Campus Activities Dataset (CAD) by modeling activity attention scores using a deep CNN model. In the second phase, we use the generated attention scores for each activity category to extract significant video frames. Finally, we evaluate the inter-frame similarity index used to reduce the number of redundant frames and extract only the representative keyframes. The proposed framework is tested on different videos, and the experimental results show the performance of the proposed summarization process.

INTRODUCTION

Academic activities are carried out on university campuses as a matter of daily routine. Video cameras are installed at various points covering all these activities and stream the video footage to a backup storage server for future correspondence. During working hours the frequency of these activities are normally high, gradually lessens after the off time and reaches its minimum level in the late night. Further, these activities do not take place on weekly holidays or other public and national holidays except those explicitly scheduled. The campus's video surveillance system (VSS) continuously transmits video

Corresponding author
Imran Ahmed,
imran.ahmed@imsciences.edu.pk

streams regardless of whether these activities are happening or not. This limits the preservation of data to short periods because a huge amount of garbage data utilizes a considerable portion of the backup storage. Further, due to the non-availability of automated content-based searching mechanisms, looking manually for a particular event or an activity in such a vast video repository is a major commitment of time and resources *Choi, Chan & Yue (2016)*.

Video summarization (VS) has been an essential tool for many video analytic tasks, which produces a brief and concise representation of the visual content of a video. A qualitative video summary has two essential features. First, it must be representative, such that it covers all the important events of the original video, and second, it should have minimal redundancy (*Ejaz, Tariq & Baik, 2012*). Several VS methods have been proposed for various application domains like electronic media such as sports, TV series, movies, documentaries, *etc.*, personal videos such as lifelogging, birthday parties, medical videos such as digital endoscopies, online databases such as YouTube, and many surveillance applications such as homes, traffic, agriculture, industries, public places and so on.

As far as academic campuses are concerned, a negligible amount of research work has been done in diverse and demanding areas. In this paper, we propose a deep learning-based fine-grain activities recognition and summarization method for the long video surveillance streams covering, for instance, two major academic activities Examination and Class Lecture. The proposed method will allow users to quickly navigate through large video databases and take faster decisions regarding selection, deletion, sharing, and consuming video content. The major contribution of this research work includes the introduction of a trained model for the above-stated academic activities and activity-based summarization method in a novel campus domain.

The rest of the paper is organized as follows: "Related Work" contains a review of existing video summarization techniques proposed by researchers. In "Methodology", we explain our dataset and activity-based summarization framework. "Experimental Results and Discussion" presents our experimental results, and finally, "Conclusion" concludes our work and provides some future directions.

RELATED WORK

Researchers have proposed solutions for various video summarization problems in a particular application domain. However, these methods vary considerably in terms of efficiency and complexity, depending on the nature of the problem at hand.

Early methods of crucial frame selection for summary generation uses low-level feature like color, motion, and texture analysis. Obtaining color histogram (*Ejaz, Tariq & Baik, 2012*) for each of the successive frames of a video and selecting keyframes based on histogram difference has been one of the most commonly used approaches with the color-based methods. Similarly, optical flow analysis (*Deng & Manjunath, 1997*), models motion patterns to evaluate motion histograms and wavelet transforms, for texture analysis (*Mahmoud, Ismail & Ghanem, 2013*), are among the most popular approaches for evaluating frame saliency to generate video summaries. Instead of using a single feature, researchers also proposed unified frameworks (*Gianluigi & Raimondo, 2006*) which make

use of multiple low-level features to generate more valuable summaries. However, the importance of these features greatly varies for different types of VS problems. To overcome these discrepancies, weighted frameworks (*Ioannidis, Chasanis & Likas, 2016*) were proposed, which assign weights to different features for practical saliency calculation specific to the problem under investigation. With the advancements in video technology, the use of low-level features has no longer remained an ideal approach for VS. High-level features like face detection (*Lee et al., 2015*), objects identification (*Lin et al., 2017*), emotions (*Lan et al., 2016*) and gestures detection (*Dammak, Wali & Alimi, 2015*) produces more appealing video summaries for high quality and complex video contents.

Video summarization methods greatly vary depending on the nature of the problem in various application domains and the content of the video. Clustering is one of the most extensively used techniques for summarization, where the visual contents share resemblance over different time periods in the video. *Zhuang et al. (1998)* authors have proposed a VS method by grouping video frames that potentially have high similarity, calculating cluster center in each group, and finally combining all the cluster centers to generate a static video summary. Several approaches such as Partitioned Clustering (*Valdés & Martnez, 2007*), Spectral Clustering (*Damnjanovic et al., 2008*; *Stefanidis et al., 2000*) and K-Mean clustering (*Amiri & Fathy, 2010*) have been used for different summarization problems. Many techniques such as High-Density Peak Search (*Wu et al., 2017*), Self-Organizing Maps, and Gaussian Mean (*John, Nair & Kumar, 2017*) are proposed by researchers to improve the keyframe selection process further to produce less redundant and qualitative video summaries.

Professional videos normally have a hierarchical structure and consist of a finite number of scenes where a scene is a collection of multiple shots. Shot boundary detection in such videos is a complex task. Transformation Co-efficient distance (*Girgensohn & Foote, 1999*), Frame packing algorithm (*Uchihachi, Foote & Wilcox, 2003*), Hierarchical clustering using Color Histogram distance (*Chheng, 2007*) and SIFT point distribution histogram (*Hannane et al., 2016*) are among the most commonly used techniques proposed by different researchers for detecting shot boundaries. Salient frames from each shot are combined to obtain a summary of the full-length video. As opposed to professional videos, the contents of surveillance videos are highly unstructured, normally having a static background and involve very less or even no camera motion. In such situations, the most desirable aspect of performing the intended VS is the detection of the behavior of the moving objects. *Qiu et al. (2008)*, and *Stefanidis et al. (2000)* authors extract and use trajectory information of the moving nodes using Spatio-temporal features for the generation of video summary. Similarly, Curve Simplification based summarization (*Bulut & Capin, 2007*; *Song et al., 2016*) evaluate frame saliency by analyzing motion behaviors more efficiently in a situation where camera motion is involved in surveillance videos.

Apart from normal video summarization, which focuses on the evaluation of individual frame saliency, the detection of events of interest has been a challenging task in event-based video summarization (*Wang & Ngo, 2011*). *Damnjanovic et al. (2008)* authors use a reference frame representing an activity to identify frames with identical activity by

calculating frame difference, assigning weights to each different frame, and finally selecting high weight frames as part of the summary. A Context-aware video summarization (*Zhang, Zhu & Roy-Chowdhury, 2016*) process performs Spatio-temporal correlation among events and uses this information for summary generation in surveillance videos. Another event detection method (*Zhang et al., 2017*) where the proposed algorithm identifies a common event along with multiple videos and performs saliency calculation in an unsupervised manner. The performance of event-based summarization significantly depends on selecting and extracting appropriate features for classification. Rather than using hand-crafted features, researchers recommend learned features because Convolution Neural Network (CNN) inherits distinctive learning ability for classification of video contents such as complex activity classification tasks (*Karpathy et al., 2014*; *Rodrguez-Moreno et al., 2019*).

Poleg et al. (2016), *Alom et al. (2017)*, and *Ji et al. (2012)* achieved state-of-the-art results while using CNN features for the proposed activity detection problems. Due to its remarkable results, many deep learning based frameworks (*Ahmed et al., 2020a*; *Ahmad, Ahmed & Jeon, 2021*; *Ahmad et al., 2020*, *2020b*; *Mahasseni, Lam & Todorovic, 2017*; *Plummer, Brown & Lazebnik, 2017*) are proposed by the researchers for video summarization in various application domains. Authors performed summarization of surveillance videos (*Muhammad, Hussain & Baik, 2020*), egocentric videos (*Jain, Rameshan & Nigam, 2017*; *Del Molino et al., 2016*) and industrial videos (*Muhammad et al., 2019*) using deep features. Similarly, *Koutras, Zlatinsi & Maragos (2018)* presented an event based method using a combination of 3D-CNN and 2D-CNN architectures, a unified framework, for high quality structured videos. Deep learning based methods (*Ahmed et al., 2021b*; *Ahmed, Anisetti & Jeon, 2021c*; *Ahmed et al., 2021d*; *Ahmed & Jeon, 2021*; *Ahmed et al., 2021a*; *Ahmed et al., 2022*), is also used by researcher such as *Wasim et al. (2021)* presented a novel deep learning based automated system for academic activities recognition.

Researchers have introduced many state-of-art activity recognition methods for various application domains with the availability of several activity data sets. Unfortunately, none of these data sets or methods have been developed explicitly for academia and do not cover academic activities. Thus, an automated deep learning-based academic activities recognition system is introduced in this work.

METHODOLOGY

Our proposed Campus Surveillance Video Summarization (CSVS) method consists of five main tasks, as shown in Fig. 1. The first one is the development of a realistic academic activities dataset, because the existing datasets (*Fan, Wang & Huang, 2016*; *Singh, Velastin & Ragheb, 2010*; *Denina et al., 2011*; *Leyva, Sanchez & Li, 2017*; *Wang et al., 2014*; *Malon et al., 2018*; *Awad et al., 2021*) researchers have been using to solve various summarization problems, lack academic activities. Second, training a CNN model with a novel set of academic activities. The third step performs activity recognition using activity attention scores generated by the trained model. In the next phase, saliency evaluation and significant activity frame extraction are performed to obtain a coarse-grained video
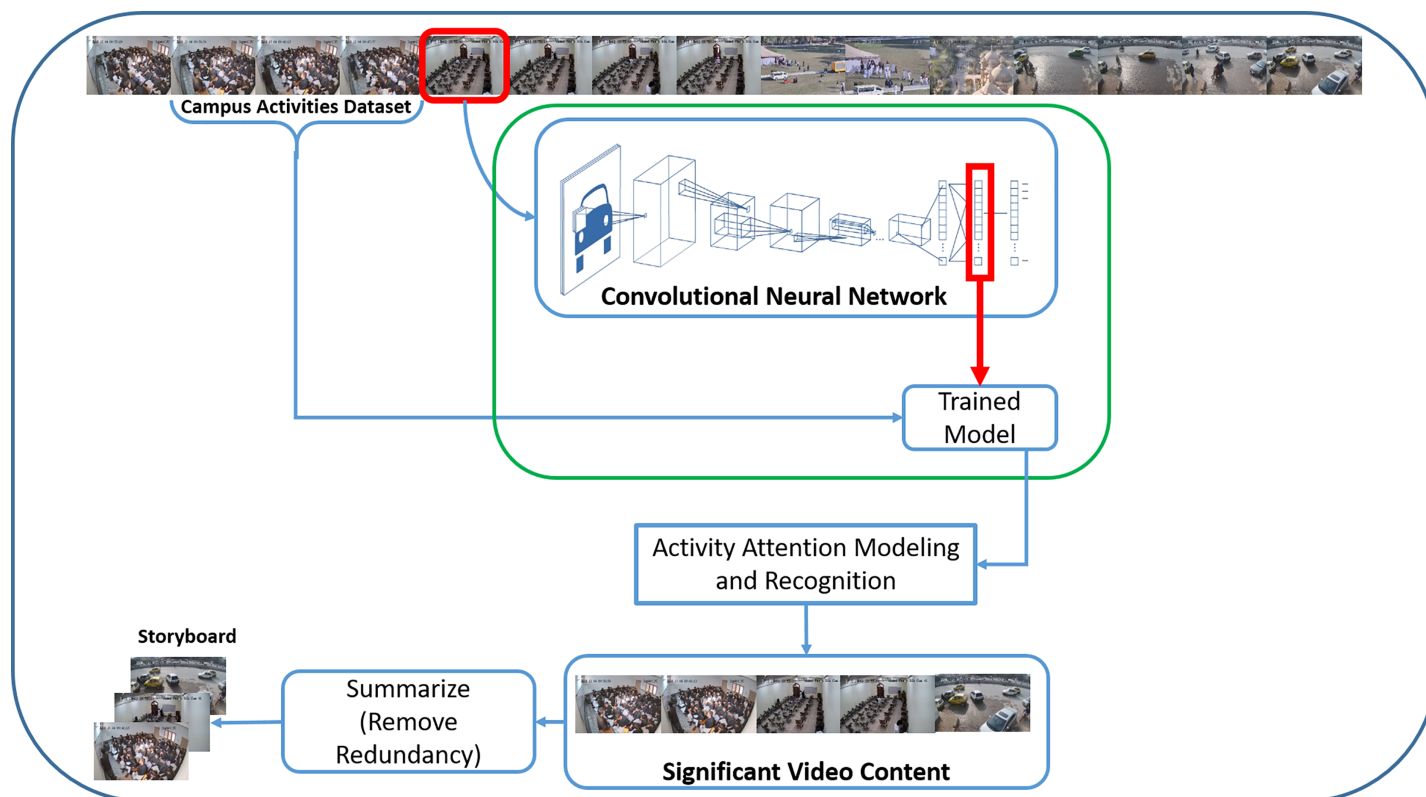
Figure 1 CNN based campus surveillance video summarization framework.

summary based on activity attention score. Finally, the removal of redundant frames and extraction of keyframes is carried out to generate the final summary.

Data set

The video data required for this research work is obtained from the surveillance repository of Islamia College Peshawar (ICP), a public sector university situated in Khyber Pakhtunkhwa, providing higher education under the governance of Higher Education Commission (HEC), Pakistan. The ICP smart campus (https://icp.edu.pk/page.php?abc=201905090555484) surveillance network consisting of 771-night vision cameras, installed at all the sensitive areas, including the classrooms, capture, and streams all the routine activities to a 92 Terabyte data server.

For instance, our proposed Campus Activities Dataset (CAD) covers two main academic activities that generally happen within a classroom, Class and Examination (Exam). Class refers to an ongoing teaching session, while Exam is some written assessment conducted from the students. Both these activities are further categorized into a total of 14 fine-grain activities. The Class is divided into six categories: Writing on Board, Explanation (Teacher explaining a topic), Board Cleaning, Class Entry, Class Exit, and Class Break. Entry and Exit refer to somebody entering or leaving the classroom while Class Break is the recess period when students are informally sitting in the classroom

Table 1 Characteristics of campus activities dataset.

S. No.	Category	No. of video samples	Duration (H:M:S)	No. of frames
1	Writing on Board (WB)	104	00:28:52	14,507
2	Explanation (Exp)	161	00:39:27	15,052
3	Board Cleaning (BC)	63	00:16:11	9,237
4	Class Entry (CE)	136	00:19:03	15,271
5	Class Exit (CEx)	96	00:13:09	11,918
6	Class Break (CB)	42	00:09:36	6,634
7	Paper Distribution (PD)	86	00:21:43	14,194
8	Paper Attempt (PA)	121	00:18:50	12,871
9	Paper Signature (PS)	87	00:22:49	15,227
10	Attendance (Att)	109	00:23:36	11,829
11	Paper Collection (PC)	207	00:26:19	16,102
12	Exam Entry (EE)	121	00:15:17	15,363
13	Exam Exit (EEx)	131	00:15:59	13,488
14	Exam Break (EB)	52	00:11:42	6,357

during an off session. Similarly, Exam is divided into eight categories which are Paper Distribution, Paper Attempt (students writing their answers), Paper Signature (Invigilator signing the answer book), Attendance (Ensure student presence during exam), Paper Collection, Exam Entry, Exam Exit, and Exam Break. This categorization aims to cover the whole activity for the proposed activity recognition and summarization task. The dataset development is done in the following steps.

- Sufficient video streams are collected from the surveillance data server. These videos are then scrutinized to select appropriate streams based on different camera angles, camera location, visual clarity, and the required contents.
- For each of these 14 activities, appropriate video segments are manually cropped from the selected video streams. The video segments are variable in length depending on the duration of a particular activity in the video. The number of video segments in each category is shown in Table 1.
- The video segments are converted into video frames for more precision. All the irrelevant, noisy, and redundant frames are removed from each category, and we are left with only those frames that best represent the target activity, Table 2.
- Finally, the frames are labeled in sequential order as they appear in the original video segment.

Model training

As discussed in "Related Work", CNN models possess significant learning ability and have a high accuracy rate in activity classification on large datasets. Several CNN models (*Dhillon & Verma, 2020*) exist that the researchers have been using for various problems. While using a pre-trained CNN model like AlexNet or VGG-16, having a lengthy

Table 2 Architecture of the CNN model.

Layer	Kernels	Size	Stride	Activation
Conv – 1	128	(3,3)	(1,1)	Relu
Max-Pooling	–	(2,2)	(2,2)	–
Conv – 2	64	(3,3)	(1,1)	Relu
Max-Pooling	–	(2,2)	(2,2)	–
Conv – 3	32	(3,3)	(1,1)	Relu
Max-Pooling	–	(2,2)	(2,2)	–
Conv – 4	16	(3,3)	(1,1)	Relu
Max-Pooling	–	(2,2)	(2,2)	–
FC-01	–	500	–	–
FC-02	–	1000	–	–
Output	–	14	–	Softmax

architectural pipeline and a large number of kernels, for the proposed problem would encounter additional computational and processing complexity because there is significantly less variation among the video frames in the CAD dataset, the cameras are fixed, and no camera motion is involved. Hence, we use a customized model which consists of 04 convolution layers, 04 max-pooling layers, 02 fully-connected (FC) layers, and 01 classification layer in such a way that a max-pool layer follows each convolution layer. Next to the last max-pool, two FC layers are followed by a Softmax classification layer, as shown in Table 2.

Before starting the training process, the CAD dataset is initially split into Training, and Test sets where the Training set contains 90% and the Test set contains 10% of the total frames in each category. The training data is further divided into Train and Validation sets in such a way that each set comprises 70% and 30% frames, respectively. The video frames (VF) are resized to dimensions (224 × 224) having 03 color channels. The input layer read each video frame and the convolution function, Eq. (1), produces an output feature map of size ($h1, w1, c2$), Eq. (2), by applying a filter kernel K of size (k, k, c) to each VF with dimensions (h, w, c).

$$Convolve(VF, K)(x, y) = \sum_{(i=1)}^{h} \sum_{(j=1)}^{w} \sum_{(k=1)}^{c} K_{(i,j,k)} VF_{(x+i-1,y+j-1,k)} \tag{1}$$

$$F_{(x,y)}(Convolve(VF, K)) = \left(\frac{h + 2P - k}{S} + 1\right), \left(\frac{w + 2P - k}{S} + 1\right) \; for \; S > 0 \tag{2}$$

or

$$F_{(x,y)} = (h + 2P - k, w + 2P - k) \; for \; S = 0 \tag{3}$$

where,

h = height of video frame

w = width of the video frame

c = no. of color channels

k = size the kernel

P = Padding size, and

S = Stride

The max-pool layer reduces the number of parameters received from the convolution layer by keeping only the high-level features representation and thus avoiding computation of passive weights, which ultimately helps reduce the chances of over-fitting. This operation is carried out by applying a function, Eq. (3), which only reduces the dimensions of the frame $F_{(x,y)}$ while keeping the number of channels, c, unchanged.

$$x, y(Pool(F_{(x,y)})) = \frac{(h + 2P - k)}{S + 1}, \frac{(w + 2P - k)}{S + 1}, c) \ for \ S > 0 \tag{4}$$

or

$$x, y(Pool(F_{(x,y)})) = (h + 2P - k, w + 2P - k, c) \ for \ S = 0 \tag{5}$$

The feature weights, f_w, are converted to a one-dimensional vector by a flatten layer, passes it to the fully connected layer, which generates another vector of learned parameters z_j, Eq. (4), and mapping each parameter to every node j of the ith FC layer.

$$z_j^{[i]} = \sum_{(l=1)}^{n_{i-1}} fw_j, l^{[i]}, a_l^{[i-1]} + b_j^{[i]} \rightarrow a_j^{[i]} = \phi^{[i]}\left(z_j^{[i]}\right) \tag{6}$$

where, $a_l^{[i-1]}$ is the result of the last pool layer and $b_j^{[i]}$ is the bias. Finally, a total of $n_{l-1} \times n_l$ learned weights $w_{j,l}$ at lth layer are obtained.

Activity recognition and attention modeling

In this step, we use our trained model for the proposed activity recognition task. Initially, the target video Vt is first converted into a set of NF frames, Eq. (5); then, features are extracted for each of the input frames using Eq. (2) and stored as a one-dimensional vector FW.

$$V_t = \{F(t + i) | i = 0, 1, 2 \dots n_{NF-1}\} \tag{7}$$

The activity attention function, Eq. (6), calculates attention score in the range (0, 1) by obtaining exponential value for each element of FW and dividing it with the sum of all values. The output layer, softmax, is configured with 14 categories, and for each of the output categories, an attention score is calculated for every individual frame in the video sequence, *i.e.*, a 1×14 vector is maintained for every video frame. Thus we obtain a total of NF vectors of the size 1×14 for all the frames of V_t such that each $(1, i)$th value, for $i = 1, 2, \dots, 14$, corresponds to one particular output category.

$$S(FW)i = \frac{e^{FW_i}}{\sum_{j=1}^{k} e^{FW_j}} \ for \ i = 1 \ K \ and \ FW = (z_1 z_k) \tag{8}$$

The activity recognition process is carried out by extracting activity frames based on their generated attention scores for the corresponding activity category. The decision that what activity is going within the input video frame is indicated by the highest attention value associated with a particular category.

Saliency evaluation and summary generation

The attention score, representing the confidence level of each of the activity frames, may be different from others depending on the content of the video frame. For summary generation, we extract the salient frames containing significant visual content. An attention score of 1, *i.e.*, 100% match, associated with any activity frame is considered a significant frame, and all other frames with lower values will be dropped. The obtained sequence of frames belonging to the same activity possesses a high level of inter-frame similarity. We remove all such redundant frames and extract only the keyframes (KF) with a high content variation. We apply the Mean Square Deviation (MSD) function for keyframe extraction, which returns a difference value of the two activity frames. For any two frames, F1 and F2, with dimensions ($m \times n$), the difference value is evaluated as follows.

$$DiffVal = \frac{1}{m \times n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [F1(i,j), F2(i,j)]^2 \tag{9}$$

Using Eq. (7), difference values are evaluated for the whole frame sequence, and a one-dimensional vector is maintained where each index location contains the difference value of the two consecutive frames. In this way, a total of n difference values are obtained for the N number of frames. A threshold value is set, Eq. (8), whereas any frame's difference value exceeding this threshold will be considered a keyframe and become part of the final summary.

$$Thresh = \frac{1}{n} \sum_{i=1}^{n} DiffVal \tag{10}$$

The threshold value is inversely proportional to the number of keyframes and can be set, high or low, to obtain the desired number of keyframes. Finally, the summary is generated by combining the keyframes into a storyboard V_{sum} as shown in Eq. (9)

$$V_s um = \sum_{i=0}^{n_K F} F_{KF}(t+i) \tag{11}$$

EXPERIMENTAL RESULTS AND DISCUSSION

All the experiments are performed on a machine with Intel Core-i5 3.2 GHz CPU, 24 GB RAM, and a Graphics Processing Unit (GPU) NVidia GeForce GTX 1080 Ti. The programming environment is Python with Keras, Numpy, Tensorflow, and other libraries
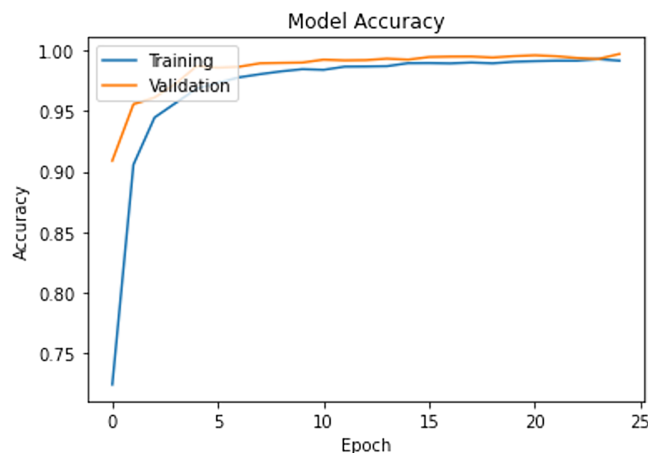
Figure 2 Training progress and train-validation accuracy.

for visualization and performance monitoring such as Scikit-Learn, OpenCV, *etc.* The programming environment is configured on Windows 10 operating system.

The model is trained on the CAD dataset with 14 categories. The training process is set for 25 epochs with a learning rate of 0.01 and a batch size of 32. Adam optimizer is used to handle noisy content as surveillance data are most vulnerable to various noises. The validation data is used to check the model's learning progress during the training process. Figure 2 shows the progress of the training process, and it is clear that the model presents smooth generalization on our dataset while achieving almost 98% Training-Validation accuracy.

The trained model is further evaluated on the test data, which was 10% of the CAD dataset as discussed in "Model Training", and it can be seen in Fig. 3 that it has correctly predicted all the test frames.

The accuracy and various other classification metrics like Precision, Recall and F-Measure confirm the efficiency of our model and are depicted in Fig. 4.

We have obtained fresh videos from our surveillance repository for activity recognition and cropped appropriate segments from these videos. The video segments were then integrated into a single video in such a way that each video was representative with respect to the diversity of contents, camera angles, and camera location. The proposed framework was tested with four (04) videos, Video 1, Video 2, Video 3, and Video 4. Video 1 includes the set of all activities that our model was supposed to recognize. Video 2 contains a sequence of video segments that were taken from examination videos, and Video 3 is a collection of teaching video segments obtained from classroom lecture videos. In addition, however, manually recorded selective academic video segments from YouTube videos were combined into Video 4.

Video 1 is supplied to the trained model for evaluation. For each frame of this video, the model has successfully generated an attention score. Figure 5 shows the attention scores generated for all of the activities present in Video 1.
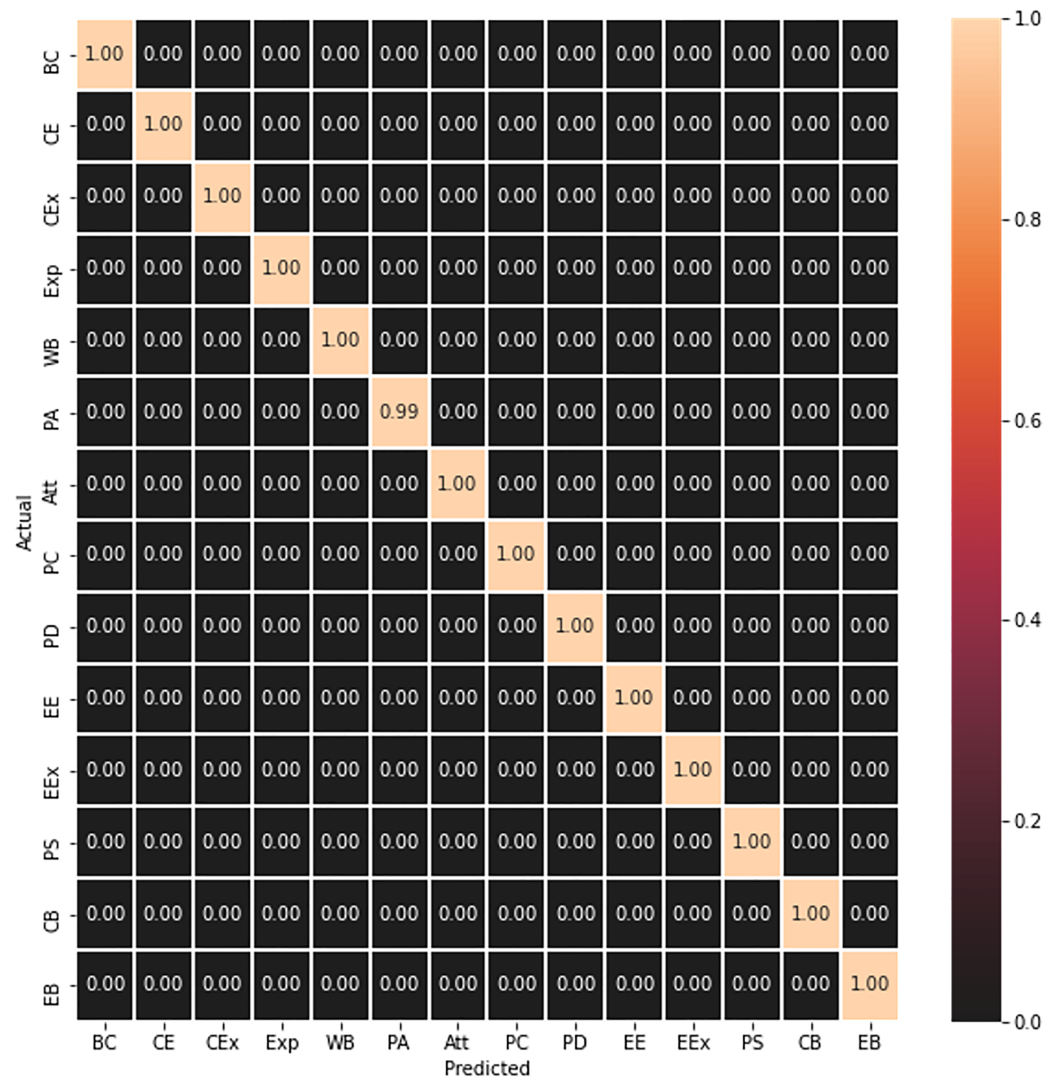
Figure 3 Concept matrix representing prediction result of test frames.

The video frames representing different activities, such as the CAD dataset, were extracted by obtaining the attention scores associated with each category, as shown in Fig. 6. The majority of the activity frames have higher attention values and are significantly closer to the highest value, *i.e.*, 1. It is because these activities are parts of one of the main activities, *i.e.*, Class Lecture or Exam. For example, during an ongoing lecture session, a teacher explains a topic, writes on the board, an entry or exit to or from the classroom happens during the same session. Therefore, the activities are differentiated on the basis of the highest match, maximum attention value, for a particular category. Similarly, the same process has been adopted for activities that belong to the examination.

```
                        Classification Report
                        precision    recall  f1-score

        Board Cleaning       1.00      1.00      1.00
           Class Entry       1.00      1.00      1.00
            Class Exit       1.00      1.00      1.00
          Explaination       0.99      1.00      1.00
       Writing on Board      1.00      1.00      1.00
         Paper Attempt       1.00      0.99      1.00
            Attendence       1.00      1.00      1.00
      Paper Collection       1.00      1.00      1.00
     Paper Distribution      1.00      1.00      1.00
            Exam Entry       1.00      1.00      1.00
             Exam Exit       1.00      1.00      1.00
        Paper Signature      1.00      1.00      1.00
           Class Break       1.00      1.00      1.00
            Exam Break       1.00      1.00      1.00

              accuracy                           1.00
             macro avg       1.00      1.00      1.00
          weighted avg       1.00      1.00      1.00
```

Figure 4 Various classification metrics showing the classification performance of the trained model.
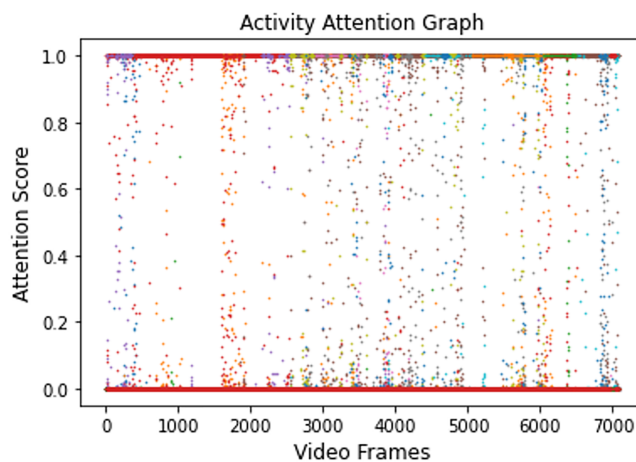
Figure 5 Activity attention plot showing all attention values of Video 1.

Video 1 is summarized by obtaining all the salient frames from all categories. This is achieved by obtaining the highest attention value associated with each frame. Figure 7 shows the saliency graph of Video 1.

As discussed in Section 3.4, a sequence of 6,660 significant frames was extracted from a total of 7,078 video frames with a percentage accuracy, measured as a ratio of significant frames to the total number of video frames of 94%. The redundant frames were removed from the sequence of significant frames by applying Eq. (7), and the final storyboard contains a total of 295 keyframes, based on Eq. (5), which is approximately 4%
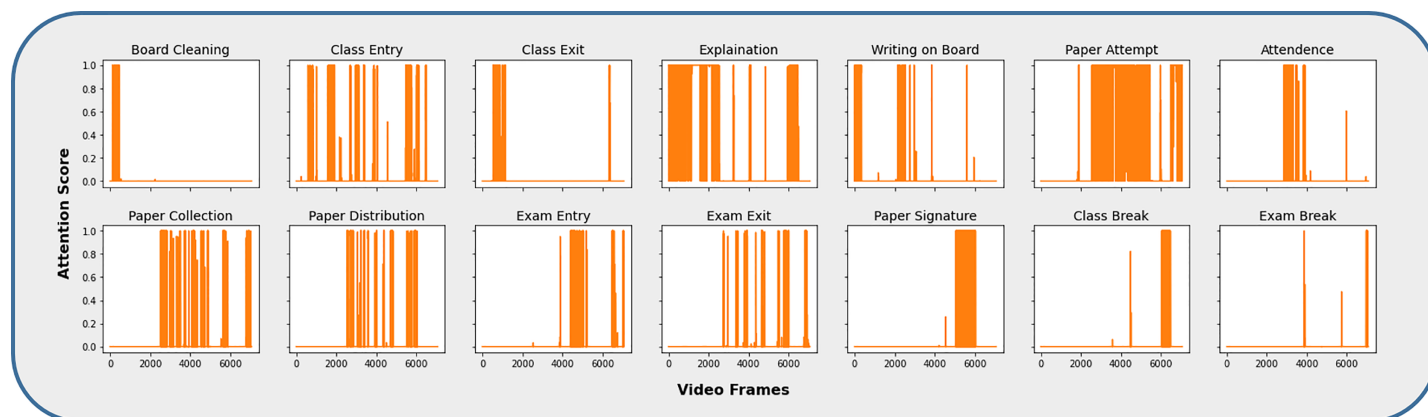
Figure 6 Video frames corresponding to individual activity category.
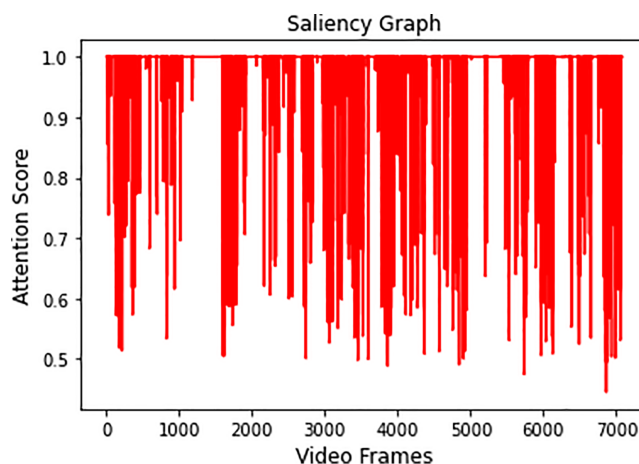
Figure 7 Graph showing salient activity frames from all the recognized activities.

of the original video containing representative frames from all the categories. The same process is repeated for Video 2, Video 3, and Video 4, where the Activity Attention and Saliency Graph of these videos are shown in Figs. 8A–8C, respectively.

The performance of the proposed framework on the four test videos has been summarized in Table 3. A quantitative summary of these four test videos is presented in Table 3. It clearly shows that for surveillance videos, Video 1, Video 2, and Video 3, the accuracy of the proposed framework insignificant activity detection is more than 94% while producing a concise summary, 4%, 1% and 3.8% respectively, of the original video. On the other hand, the model has identified only 66% of the activity frames in Video 4. It is because these videos segments are highly structured and recorded with high-resolution cameras with a limited focus in a completely different environment. However, the summarization module has efficiently produced a short summary of the identified activities covering almost all of the significant activities. Some of the representative
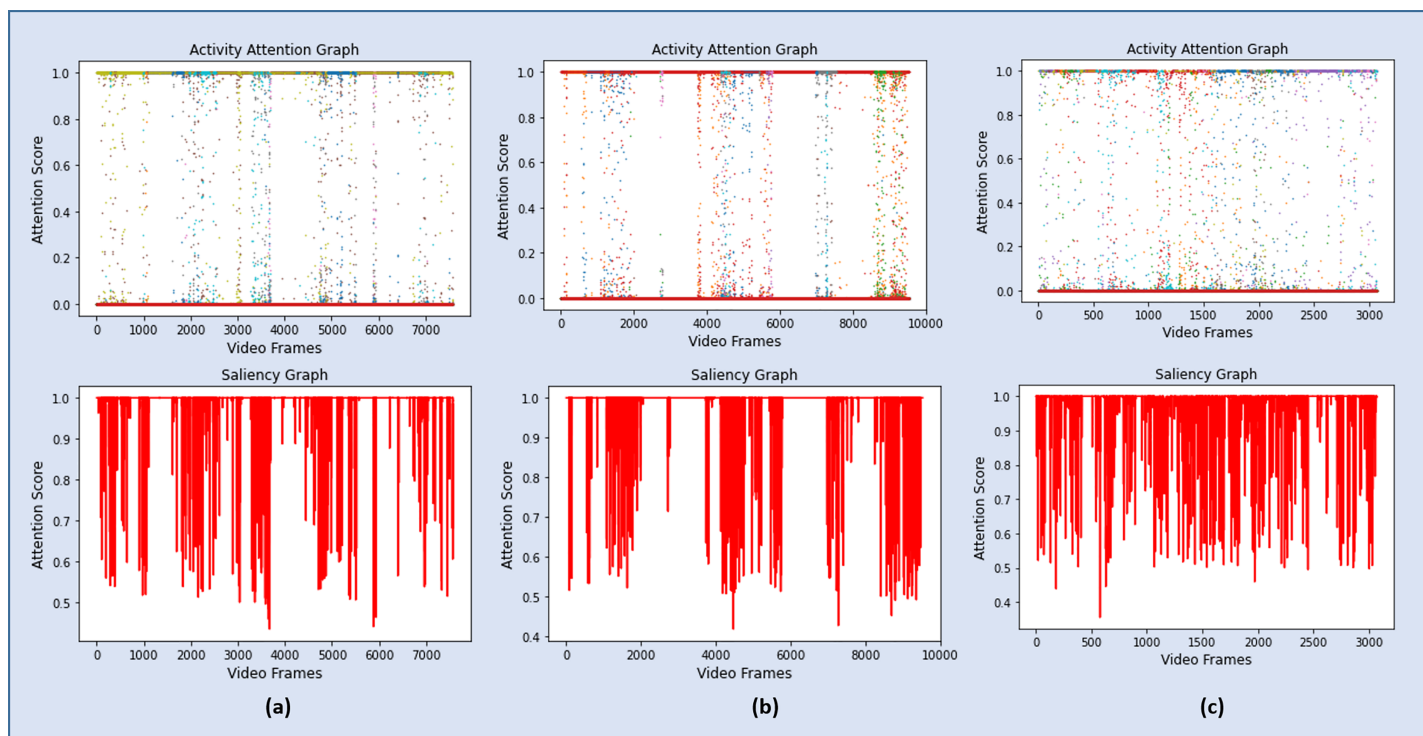
Figure 8 Graphs showing the attention scores and the corresponding salient frames for Video 2, Video 3 and Video 4.

Table 3 Performance of the proposed summarization process on test videos.

Test video	Total frames	Significant frames	Detection accuracy (%)	Key frames	Size of video (%)
Video 1	7,078	6,660	94.1	295	4
Video 2	7,570	7,193	95.0	82	1
Video 3	9,539	8,991	94.2	363	3.8
Video 4	3,070	2,031	66.1	215	7

keyframes from Video 1 are shown in Fig. 9, while Fig. 10 contains a few keyframes from the rest of the video summaries.

The keyframes were obtained by using an average threshold value, using Eq. (8), but can be set to a higher or a lower value to obtain the desired number of keyframes. One limitation associated with threshold setting is that if a significantly higher value is set, a much shorter summary can be obtained, but, in this case, one may miss many important activities and hence may compromise the information covered in the final storyboard. Similarly, a smaller value will increase the size of the final summary by the inclusion of a large number of redundant frames.

Figure 9 A representative key frame from each activity category in Video 1 story board.

Figure 10 Sample key frames extracted from Video 2, Video 3 and Video 4.

CONCLUSION

The proposed framework is mainly focused on the practical use of the video summarization process in a novel campus domain. The results state that the proposed method is producing short summaries of the target academic activities captured by the campus surveillance system. The short representation of long video streams is of significant importance because the generated summary presents a glimpse of all the important activities in a short period. Furthermore, it will allow the administrators and other stakeholders to make a faster decision regarding the selection of desired content in an efficient manner by reducing the time required in the overall navigation and video retrieval process.

One of the main hurdles in this research domain is that organizations are reluctant to provide video data due to certain constraints regarding data privacy which amounts to the scarcity of research focus in such a demanding application domain. This piece of research effort is a step toward opening ways for quality research in this area. The video analytic researcher shall primarily focus on proposing generalized trained models and unfold challenges for similar problems. We further aim to propose frameworks for generating personalized video summaries using Long Short Term Memory (LSTM) and 3D-CNN architecture.

Funding

This work was supported by the Princess Nourah bint Abdulrahman University Researchers Supporting Project number PNURSP2022R161, Princes Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia: PNURSP2022R161.

Competing Interests

Imran Ahmed is an Academic Editor for PeerJ.

Author Contributions

- Wasim Muhammad conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Imran Ahmed conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Jamil Ahmad conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

- Muhammad Nawaz performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Eatedal Alabdulkreem analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, arrange funding, and approved the final draft.
- Yazeed Ghadi analyzed the data, performed the computation work, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The code and data are available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.911#supplemental-information.

REFERENCES

Ahmad M, Ahmed I, Jeon G. 2021. An IoT-enabled real-time overhead view person detection system based on cascade-RCNN and transfer learning. *Journal of Real-Time Image Processing* **18(4)**:1129–1139 DOI 10.1007/s11554-021-01103-0.

Ahmad M, Ahmed I, Khan FA, Qayum F, Aljuaid H. 2020. Convolutional neural network-based person tracking using overhead views. *International Journal of Distributed Sensor Networks* **16(6)**:1550147720934738 DOI 10.1177/1550147720934738.

Ahmed I, Ahmad M, Ahmad A, Jeon G. 2021a. Top view multiple people tracking by detection using deep sort and YOLOv3 with transfer learning: within 5g infrastructure. *International Journal of Machine Learning and Cybernetics* **12(11)**:3053–3067 DOI 10.1007/s13042-020-01220-5.

Ahmed I, Ahmad M, Khan FA, Asif M. 2020a. Comparison of deep-learning-based segmentation models: using top view person images. *IEEE Access* **8**:136361–136373 DOI 10.1109/ACCESS.2020.3011406.

Ahmed I, Ahmad M, Rodrigues JJ, Jeon G. 2021b. Edge computing-based person detection system for top view surveillance: using centernet with transfer learning. *Applied Soft Computing* **107(3)**:107489 DOI 10.1016/j.asoc.2021.107489.

Ahmed I, Anisetti M, Jeon G. 2021c. An IoT-based human detection system for complex industrial environment with deep learning architectures and transfer learning. *International Journal of Intelligent Systems* **10(3)**:91 DOI 10.1002/int.22472.

Ahmed I, Camacho D, Jeon G, Piccialli F. 2022. Internet of health things driven deep learning-based system for non-invasive patient discomfort detection using time frame rules and pairwise keypoints distance feature. *Sustainable Cities and Society* **79(3)**:103672 DOI 10.1016/j.scs.2022.103672.

Ahmed I, Din S, Jeon G, Piccialli F, Fortino G. 2020b. Towards collaborative robotics in top view surveillance: a framework for multiple object tracking by detection using deep learning. *IEEE/CAA Journal of Automatica Sinica* **8(7)**:1253–1270 DOI 10.1109/JAS.2020.1003453.

Ahmed I, Jeon G. 2021. A real-time person tracking system based on SiamMask network for intelligent video surveillance. *Journal of Real-Time Image Processing* **18(5)**:1803–1814 DOI 10.1007/s11554-021-01144-5.

Ahmed I, Jeon G, Chehri A, Hassan MM. 2021d. Adapting Gaussian YOLOv3 with transfer learning for overhead view human detection in smart cities and societies. *Sustainable Cities and Society* **70**:102908 DOI 10.1016/j.scs.2021.102908.

Alom MZ, Alam M, Taha TM, Iftekharuddin KM. 2017. Object recognition using cellular simultaneous recurrent networks and convolutional neural network. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. Piscataway: IEEE, 2873–2880.

Amiri A, Fathy M. 2010. Hierarchical keyframe-based video summarization using QR-decomposition and modified-means clustering. *EURASIP Journal on Advances in Signal Processing* **2010(1)**:1–16 DOI 10.1155/2010/892124.

Awad G, Butt AA, Curtis K, Fiscus J, Godil A, Lee Y, Delgado A, Zhang J, Godard E, Chocot B, Diduch L, Liu J, Smeaton AF, Graham Y, Jones GJF, Kraaij W, Quenot G. 2021. Trecvid 2020: a comprehensive campaign for evaluating video retrieval tasks across multiple application domains. *ArXiv Preprint. Available at* arXiv:2104.13473.

Bulut E, Capin T. 2007. Key frame extraction from motion capture data by curve saliency. In: *Computer Animation and Social Agents*. Vol. 20.

Chheng T. 2007. *Video summarization using clustering*. Irvine: Department of Computer Science University of California.

Choi T-M, Chan HK, Yue X. 2016. Recent development in big data analytics for business operations and risk management. *IEEE Transactions on Cybernetics* **47(1)**:81–92 DOI 10.1109/TCYB.2015.2507599.

Dammak M, Wali A, Alimi AM. 2015. Viewer's affective feedback for video summarization. *Journal of Information Processing Systems* **11(1)**:76–94.

Damnjanovic U, Fernandez V, Izquierdo E, Martinez JM. 2008. Event detection and clustering for surveillance video summarization. In: *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*. Piscataway: IEEE, 63–66.

Del Molino AG, Tan C, Lim J-H, Tan A-H. 2016. Summarization of egocentric videos: a comprehensive survey. *IEEE Transactions on Human-Machine Systems* **47(1)**:65–76 DOI 10.1109/THMS.2016.2623480.

Deng Y, Manjunath B. 1997. Content-based search of video using color, texture, and motion. In: *Proceedings of International Conference on Image Processing*. Vol. 2. Piscataway: IEEE, 534–537.

Denina G, Bhanu B, Nguyen HT, Ding C, Kamal A, Ravishankar C, Roy-Chowdhury A, Ivers A, Varda B. 2011. Videoweb dataset for multi-camera activities and non-verbal communication. In: *Distributed Video Sensor Networks*. Berlin: Springer, 335–347.

Dhillon A, Verma GK. 2020. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence* **9(2)**:85–112 DOI 10.1007/s13748-019-00203-0.

Ejaz N, Tariq TB, Baik SW. 2012. Adaptive key frame extraction for video summarization using an aggregation mechanism. *Journal of Visual Communication and Image Representation* **23(7)**:1031–1040 DOI 10.1016/j.jvcir.2012.06.013.

Fan C-T, Wang Y-K, Huang C-R. 2016. Heterogeneous information fusion and visualization for a large-scale intelligent video surveillance system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **47(4)**:593–604 DOI 10.1109/TSMC.2016.2531671.

Gianluigi C, Raimondo S. 2006. An innovative algorithm for key frame extraction in video summarization. *Journal of Real-Time Image Processing* **1(1)**:69–88 DOI 10.1007/s11554-006-0001-1.

Girgensohn A, Foote J. 1999. Video classification using transform coefficients. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*. Piscataway: IEEE, 3045–3048.

Hannane R, Elboushaki A, Afdel K, Naghabhushan P, Javed M. 2016. An efficient method for video shot boundary detection and keyframe extraction using sift-point distribution histogram. *International Journal of Multimedia Information Retrieval* **5(2)**:89–104 DOI 10.1007/s13735-016-0095-6.

Ioannidis A, Chasanis V, Likas A. 2016. Weighted multi-view key-frame extraction. *Pattern Recognition Letters* **72(1)**:52–61 DOI 10.1016/j.patrec.2016.01.027.

Jain S, Rameshan RM, Nigam A. 2017. Object triggered egocentric video summarization. In: *International Conference on Computer Analysis of Images and Patterns*. Berlin: Springer, 428–439.

Ji S, Xu W, Yang M, Yu K. 2012. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35(1)**:221–231 DOI 10.1109/TPAMI.2012.59.

John AA, Nair BB, Kumar P. 2017. Application of clustering techniques for video summarization-an empirical study. In: *Computer Science On-line Conference*. Berlin: Springer, 494–506.

Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. 2014. Large-scale video classification with convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 1725–1732.

Koutras P, Zlatinsi A, Maragos P. 2018. Exploring cnn-based architectures for multimodal salient event detection in videos. In: *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. Piscataway: IEEE, 1–5.

Lan Y, Wei S, Liu R, Zhao Y. 2016. Creating video summarization from emotion perspective. In: *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. Piscataway: IEEE, 1112–1117.

Lee Y-S, Hsu C-Y, Lin P-C, Chen C-Y, Wang J-C. 2015. Video summarization based on face recognition and speaker verification. In: *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*. Piscataway: IEEE, 1821–1824.

Leyva R, Sanchez V, Li C-T. 2017. The lV dataset: a realistic surveillance video dataset for abnormal event detection. In: *2017 5th International Workshop on Biometrics and Forensics (IWBF)*. Piscataway: IEEE, 1–6.

Lin S-T, Liao Y-H, Tsao Y, Chien S-Y. 2017. Object-based on-line video summarization for internet of video things. In: *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. Piscataway: IEEE, 1–4.

Mahasseni B, Lam M, Todorovic S. 2017. Unsupervised video summarization with adversarial LSTM networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 202–211.

Mahmoud KM, Ismail MA, Ghanem NM. 2013. VSCAN: an enhanced video summarization using density-based spatial clustering. In: *International Conference on Image Analysis and Processing*. Berlin: Springer, 733–742.

Malon T, Roman-Jimenez G, Guyot P, Chambon S, Charvillat V, Crouzil A, Péninou A, Pinquier J, Sèdes F, Sénac C. 2018. Toulouse campus surveillance dataset: scenarios,

soundtracks, synchronized videos with overlapping and disjoint views. In: *Proceedings of the 9th ACM Multimedia Systems Conference*. 393–398.

Muhammad K, Hussain T, Baik SW. 2020. Efficient CNN based summarization of surveillance videos for resource-constrained devices. *Pattern Recognition Letters* **130**:370–375 DOI 10.1016/j.patrec.2018.08.003.

Muhammad K, Hussain T, Del Ser J, Palade V, De Albuquerque VHC. 2019. Deepres: a deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios. *IEEE Transactions on Industrial Informatics* **16(9)**:5938–5947 DOI 10.1109/TII.2019.2960536.

Plummer BA, Brown M, Lazebnik S. 2017. Enhancing video summarization via vision-language embedding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 5781–5789.

Poleg Y, Ephrat A, Peleg S, Arora C. 2016. Compact CNN for indexing egocentric videos. In: *2016 IEEE winter conference on applications of computer vision (WACV)*. Piscataway: IEEE, 1–9.

Qiu X, Jiang S, Liu H, Huang Q, Cao L. 2008. Spatial-temporal attention analysis for home video. In: *2008 IEEE International Conference on Multimedia and Expo*. Piscataway: IEEE, 1517–1520.

Rodrguez-Moreno I, Martnez-Otzeta JM, Sierra B, Rodriguez I, Jauregi E. 2019. Video activity recognition: State-of-the-art. *Sensors* **19(14)**:3160 DOI 10.3390/s19143160.

Singh S, Velastin SA, Ragheb H. 2010. Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods. In: *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. Piscataway: IEEE, 48–55.

Song X, Sun L, Lei J, Tao D, Yuan G, Song M. 2016. Event-based large scale surveillance video summarization. *Neurocomputing* **187(17)**:66–74 DOI 10.1016/j.neucom.2015.07.131.

Stefanidis A, Partsinevelos P, Agouris P, Doucette P. 2000. Summarizing video datasets in the spatiotemporal domain. In: *Proceedings 11th International Workshop on Database and Expert Systems Applications*. Piscataway: IEEE, 906–912.

Uchihachi S, Foote JT, Wilcox L. 2003. Automatic video summarization using a measure of shot importance and a frame-packing method. US Patent 6,535,639.

Valdés V, Martnez JM. 2007. On-line video skimming based on histogram similarity. In: *Proceedings of the International Workshop on TRECVID Video Summarization*. 94–98.

Wang F, Ngo C-W. 2011. Summarizing rushes videos by motion, object, and event understanding. *IEEE Transactions on Multimedia* **14(1)**:76–87 DOI 10.1109/TMM.2011.2165531.

Wang T, Gong S, Zhu X, Wang S. 2014. Person re-identification by video ranking. In: *European Conference on Computer Vision*. Berlin: Springer, 688–703.

Wasim M, Ahmed I, Ahmad J, Hassan MM. 2021. A novel deep learning based automated academic activities recognition in cyber-physical systems. *IEEE Access* **9**:63718–63728 DOI 10.1109/ACCESS.2021.3073890.

Wu J, Zhong SH, Jiang J, Yang Y. 2017. A novel clustering method for static video summarization. *Multimedia Tools and Applications* **76(7)**:9625–9641 DOI 10.1007/s11042-016-3569-x.

Zhang D, Han J, Jiang L, Ye S, Chang X. 2017. Revealing event saliency in unconstrained video collection. *IEEE Transactions on Image Processing* **26(4)**:1746–1758 DOI 10.1109/TIP.2017.2658957.

Zhang S, Zhu Y, Roy-Chowdhury AK. 2016. Context-aware surveillance video summarization. *IEEE Transactions on Image Processing* **25(11)**:5469–5478 DOI 10.1109/TIP.2016.2601493.

Zhuang Y, Rui Y, Huang TS, Mehrotra S. 1998. Adaptive key frame extraction using unsupervised clustering. In: *Proceedings 1998 International Conference on Image Processing, ICIP98 (Cat. No. 98cb36269)*. Piscataway: IEEE, 866–870.