Automated Deep Learning based Video Summarization Approach for Forest Fire Detection

Saeed M. Aljaberi[1], Ahmed N. Al-Masri[2]

[1] Artificial Intelligence Department, Dubai Police, Dubai, UAE

[2] American University in the Emirates, Dubai, UAE

Emails: eljabri@live.com ; ahmed.almasri@aue.ae

Abstract

Due to the exponential increase in video data, an automated examination of videos has become essential. A significant requirement is the capability of the automated video summarization process, which is helpful in vast application areas from surveillance to security. It assists in monitoring the user application with reduced memory and time. Therefore, this paper designs an automated deep learning-based video summarization approach for forest fire detection (ADLVS-FFD). The ADLVS-FFD technique aims to summarize the captured videos and detects the existence of forest fire in it. In addition, the ADLVS-FFD technique involves different subprocesses such as frame splitting, feature extraction, and classification. Besides, a merged Gaussian mixture model (MGMM) is used to extract keyframes and features. Moreover, the long short-term memory (LSTM) model is employed to detect and classify input images into normal and forest fire images. To ensure the better performance of the ADLVS-FFD technique, a comprehensive experimental validation process takes place on a benchmark video dataset. The resultant experimental validation process highlighted the supremacy of the ADLVS-FFD technique over the recent methods.

Keywords: Video summarization, Deep learning, LSTM Model, Forest fire detection, Feature extraction

1. Introduction

Given the plethora of video content on the Web, effective video summarization facilitates viewers' browsing of and navigation in large video collections, thus increasing viewers' engagement and content consumption [1]. The application domain of automatic video summarization is wide and includes (but is not limited to) the use of such technologies by media organizations (after integrating such techniques into their content management systems), to allow effective indexing, browsing, retrieval, and promotion of their media assets; and video sharing platforms, to improve viewing experience, enhance viewers' engagement and increase content consumption [2-4].

In addition, video summarization that is tailored to the requirements of particular content presentation scenarios can be used for e.g., generating trailers or teasers of movies and episodes of a TV series; presenting the highlights of an event (e.g., a sports game, a music band performance, or a public debate); and creating a video synopsis with the main activities that took place over e.g., the last 24hrs of recordings of a surveillance camera, for time-efficient progress monitoring or security purposes [5].

Fire is an anomalous activity that might rapidly create property damage and substantial injury [6]. Based on the NAPA, the US fire sectors respond to an expected 1,319,500 fires in 2017 [7] that result in 14,670 civilian fire injuries, 3,400 civilian fire fatalities, and an expected $23 billion in straight properties

DOI: 10.5281/zenodo.5515606
Received: December 19, 2020 Accepted: August 11, 2021

damage. To decrease this disaster, fire detections with no false alarms in earlier stages are critical. Consequently, many automated fire detections technology is established and extensively employed in reality. Generally, 2 general classes of technology are recognized: conventional fire detection and alarm with CVs. Conventional fire alarm technologies are depending on a heat/smoke sensor which needs closeness to activations. This sensor needs human participation for confirming a fire if an alarm takes place. Further, this system requires many equipments for providing data on the location, size, burning, and amount of the fire.

In order to beat this limitation, researchers were examining CV based methods integrated into several kinds of supplemental sensors [8-10]. This class of technology provides large observation and offers the advantages of lesser human interference with a fast response since a fire could be confirmed with no need of visiting the fire position, and provide thorough fire data like degree, location, and size. In spite of this advantage, but, few issues remain concerned the false detection, and scheme difficulty based on the different aims. Hence, researcher participated in weighty efforts for addressing this issue based on CV technologies [11-13].

Yasmin et al. [14] established a video summary architecture with a key moments based frame clustering and selection of frames for identifying useful frames. The key moments are simpler but efficient characteristics to summarize a longer video motion and shot is the unique characteristic in giving action or event in videos i.e., employed for extracting key moment of the video frame. The movement is the scenes of video frames that contain deceleration and acceleration when the key moment takes place. According to the key moment extraction, the frame of the video is separated into distinct sets with a new comparison based agglomerate clustering method.

Basavarajaiah and Sharma [15] proposed a GVSUM method to summarize videos. This method is created with the choice of KF where main scene changes occur in that videos. Each frame of the video is allocated clustering numbers according to their visualization feature and the KF is extracting while the clustering numbers of the frames change. The visualization feature of the video is extracted from a pre-trained CNN and k-means clustering is employed on this feature after that a consecutive KF technique. But, the optimal values of cluster could be selected beforehand summarize with ASW approach.

An effective VS for monitoring scheme is presented in [16] with standardized quick sort and k-means methods. The presented method includes ID number, presampling, FS, FE, clustering, extracted frame, video summarization. Firstly, the video frame is presampled using the presented TSCS method. Then, the feature is gathered by N-Kmeans model to identify an optimal candidate's frame. Next, elect the minimal distance values relied upon cluster set is the KF selections. Eventually, the videos are summarized in orderly with quick sort technique. Messaoud et al. [17] introduce a new DeepQAMVS, which enhances various criteria: (i), chronological soundness (ii) conciseness, and (iii) representative of significant query applicable event. They designed a classified method which factorizes on 3 allocations, all gathering evidence from diverse modalities, after that a pointer networks select frame to including in the summarization. Parihar et al. [18] proposed a novel method for multi-view summarizations. The presented method employs the BIRCH clustering method for the early sets of frames to dispose of the redundant and static. The study presented a novel method for shot edge detections with a frame comparison measure Dice and Jaccard.

This paper designs an automated deep learning based video summarization approach for forest fire detection (ADLVS-FFD). The ADLVS-FFD technique aims to summarize the captured videos and detects the existence of forest fire in it. In addition, the ADLVS-FFD technique involves different subprocesses such as frame splitting, feature extraction, and finally classification. Besides, a merged gaussian mixture model (MGMM) is used for the extraction of key frames and features. Moreover, long short term memory (LSTM) model is employed for the detection and classification of input images into normal and forest fire images. In order to ensure the better performance of the ADLVS-FFD technique, a comprehensive experimental validation process takes place on benchmark video dataset.

2. The Proposed ADLVS-FFD Technique

The entire work process of the proposed MGMM-LSTM model is given here. Amongst several procedures, key frame extraction with MGMM and classification with LSTM play an important part in

DOI: 10.5281/zenodo.5515606
Received: December 19, 2020 Accepted: August 11, 2021

the entire detection process. Firstly, the video sequences would be given by the inputs to the proposed method. The transformation of video to a set of frame occurs. Later, key frame is extracted from the overall amount of frame that would be beneficial for generating a video summarization. When the key frame is extracted, few identical frames would be removed. The feature from all key frames would be elected and the collection of features is spatial analysis, color autocorgram, shape, temporal analysis, intensity, dynamic texture analysis, motion. Next, LSTM models would be summoned for classifying the image and build models. After the training models get finished, testing of input images could occur. For testing the images, the input videos are transformed into a collection of frames. Later, MGMM dependent key frame extractions occur. Lastly, the input image undergoes classification either the images include the existence of fire/not.

2.1 MGMM for feature extraction

The given raw videos in the featured spaces are deliberated by a higher dimension "feature procedure." At time t, GMM treat all the features x_t as incorporation of K Gaussian component [19], that is

$$x_t = \sum_{k=1}^{K} \omega_t \eta(X_t, \mu_k, \Sigma_k) \quad (1)$$

Where Σ_k, μ_t and $\omega_k$specifies the covariance, mean, and weight matrixes of kth components correspondingly. The likelihood density functions (pdf) are represented as follows

$$\eta(x, \mu, \Sigma) = \frac{1}{(2\pi)^n |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad (2)$$

The covariance matrixes Σ are an $n \times n$matrix which need further training to estimate. To simplify the learning process, which can be given by.

$$\Sigma = \sigma^2 I \quad (3)$$

It considers all dimensions of the featured hold related variances. Although it isn't real-world, it is used as an estimate in the higher dimension area and outcomes are satisfactory. The approximation of abovementioned parameter is relatively acceptable. For all input features x_t in time t, the distance from x_t to mean μ_k per components are defined for identifying matching components. The match occurs once the distance from x_t as well as μ_k is less compared to thrice the variances σ_k, that is

$$(x_t - \mu_k)^T (x_t - \mu_k) < 3\sigma_k \quad (4)$$

The weights get updated as

$$\omega_k, t = (1 - \alpha)\omega_{k,t-1} + \propto M_{k,t} \quad (5)$$

Where α characterizes settled learning rates, $M_{k,t}$ equal to on for matching components/ zero. Another parameter is upgraded by.

$$\mu_{k,t} = (1 - \rho)\mu_{k,t-1} + \rho x_t \quad (6)$$

$$\sigma_{k,t}^2 = (1 - \rho)\sigma_{k,t-1}^2 + \rho\left(x_t - \mu_{k,t}\right)^T \left(x_t - \mu_{k,t}\right) \quad (7)$$

Whereas ρ indicates the next learning rates and

$$\rho = (\alpha \eta x_t, \mu_k, \Sigma_k) \quad (8)$$

Once none of the components matchings are recognized, then the minimum likelihood get upgraded by $\mu_k = x_t$ high variance and low weight. The MGMM is upgrade for explaining the distributions of information streaming. However, rather than employing a settled number of clusters, this technique enables new one to be built-in required as well as offer ways to integrate equal cluster in a corresponding manner. The MGMM models are used for clustering the common information. When employing MGMM models to the video summary model, an additional process of electing a descriptive from all the clusters

DOI: 10.5281/zenodo.5515606
Received: December 19, 2020 Accepted: August 11, 2021

as a key frame. In all the levels of process, the frame placed near all the clustering means is kept as an existing key frame. The frame may endure replacing while the following frames are nearest to the mean. As the cluster means aren't static, the final sets of key frames mightn't be an optimal one that would be elected while the whole datasets are located in storage areas.

2.2 LSTM based Classification

The RNN is a variety of average ANNs that are able of demonstrating consecutive information by containing repeated connections [20]. Basically, it continues the hidden state that is regarded as "memory" of preceding input. It is determined by the detail that all neurons signify an estimated function of every preceding data. An input units $\{\dots, x_{t-1}, x_t, x_{t+1}, \dots\}$ where $= (x_1, x_2, x_3, \dots, x_N)$ are associated with hidden unit $h_t = (h_1, h_2, \cdots, h_M)$ in the hidden layers, through associates determined as weight matrix W_{IH}. All hidden units are associated with the next one with recurrent associates provided as W_{HH}. All hidden units are expressed as:

$$h_t = f_H(o_t) \tag{9}$$

where:

$$o_t = W_{IH} + W_{HH} h_{t-1} + b_h \tag{10}$$

F_h implies the nonlinear function such as sigmoid/ReLU, tanh, etc, and b_H represents the biasing vectors. The hidden layers are linked as resultant layers using weight W_{HO}. Eventually, the output $y_t = (y_1, y_2, \dots, y_P)$ are determined as:

$$y_t = f_O(W_{HO} h_t + b_o) \tag{11}$$

A similar approach as hidden layers, f_O represents the activation functions and b signifies the bias vectors. While this technique continues a memory of earlier state, in fact, it undergoes from gradient reducing issue, so developing unfeasible to longterm dependency. The different kinds of RNN are named LSTM is established in 1997 that overcome this problem. The LSTM cell follows a further sophisticated approach with outline of difficult cells which utilize "forget" gate for selecting what to forgets.

The states of LSTM accepts the succeeding models:

$$\begin{aligned} i_t &= \sigma\,(W_{xi} + W_{hi} h_{t-1} + \mathrm{W}_{ci} c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} x_t + W_{hf} h_{t-1} + \mathrm{W}_{\mathrm{cf}} c_{t-1} + b_f) \\ c_t &= f_t \otimes c_{t-1} + i_t \otimes tanh\,(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \\ o_t &= \Gamma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \\ h_t &= o_t \otimes tanh(c_t) \end{aligned} \tag{12}$$

For clarifying, the subscripts relate to first of all matrices signifies (for instance, W_{hf} implies the hidden forget weighted matrixes). Also i, f, c , & o relate to input, forget, cell, and output gates vector. The LSTM networks have resembled the primary RNNs framework given by the LSTM cell.

3. Results and Discussion

The experimental validation of the ADLVS-FFD technique is validated using a benchmark dataset [21], which comprises 267 images (including 157 Normal and 110 Forest Fire). The sample set of forest fire images are shown in Fig. 1. Followed by, the results are inspected interms of precision, recall, and accuracy.

DOI: 10.5281/zenodo.5515606
Received: December 19, 2020 Accepted: August 11, 2021

Fig. 1. Sample Forest Fire Images from test videos

Table 1 depicts a detailed comparative results analysis [22] of ADLVS-FFD technique interms of different aspects. Fig. 2 examines the precision analysis of the ADLVS-FFD technique with existing approaches on the test input images. The figure showcased that the DT and NB models have tried to showcase somewhat considerable precision of 0.8489 and 0.8390 respectively. At the same time, the RB and NN models have accomplished a moderately improved precision of 0.8676 and 0.8540 respectively. Along with that, the MGMM-KSVM technique has resulted in a near optimal performance with a precision of 0.9045. However, the proposed ADLVS-FFD technique has attained superior outcome with a maximum precision of 0.9045.

Table 1 Comparison of ADLVS-FFD technique with Existing Methods

Methods	Precision	Recall	Accuracy
Proposed Model	0.9045	0.9183	0.9145
MGMM-KSVM	0.8998	0.9082	0.9056
RB Model	0.8676	0.8853	0.8723
NN Model	0.8540	0.8618	0.8578
Naïve Bayes	0.8489	0.8539	0.8514
DT Model	0.8390	0.8123	0.8235

DOI: 10.5281/zenodo.5515606
Received: December 19, 2020 Accepted: August 11, 2021

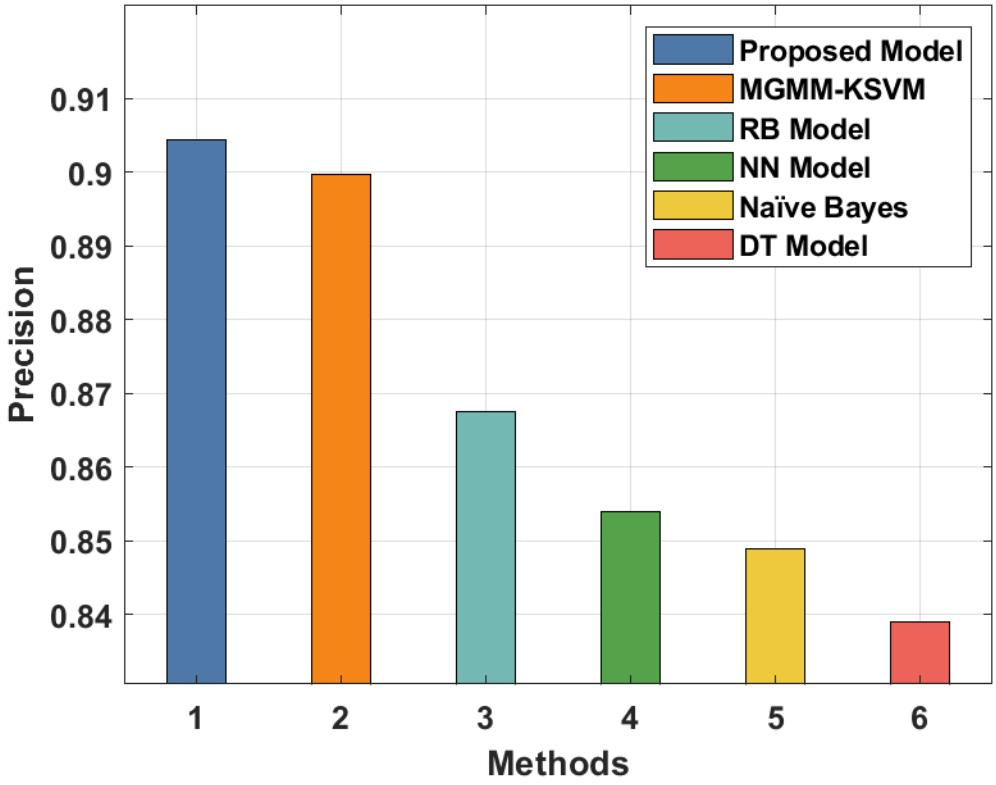

Fig. 2. Precision analysis of ADLVS-FFD with existing techniques

Fig. 3 inspects the recall analysis of the ADLVS-FFD technique with existing approaches on the test input images. The figure reported that the DT and NB models have tried to exhibit certainly improved recall of 0.8123 and 0.8539 respectively. In line with, the RB and NN models have accomplished a moderately improved recall of 0.883 and 0.8618 respectively. Simultaneously, the MGMM-KSVM technique has resulted in a near optimal performance with the recall of 0.9082. However, the proposed ADLVS-FFD technique has demonstrated improved performance with the supreme recall of 0.9183.

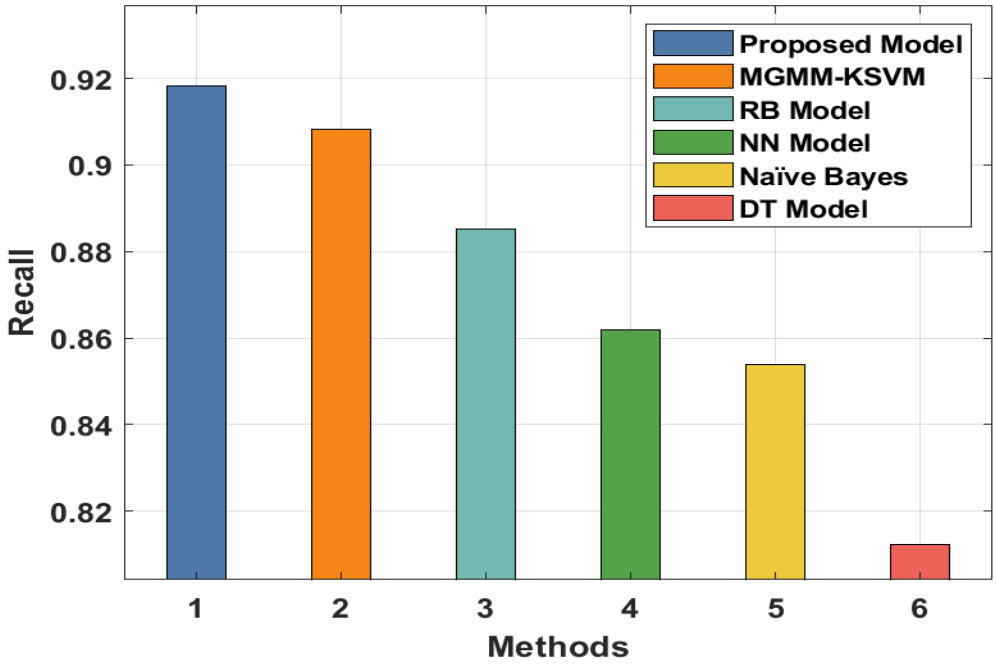

Fig. 3. Recall analysis of ADLVS-FFD with existing techniques

DOI: 10.5281/zenodo.5515606
Received: December 19, 2020 Accepted: August 11, 2021

Fig. 4 validates the accuracy analysis of the ADLVS-FFD technique with existing approaches on the test input images. The figure demonstrated that the DT and NB models have tried to showcase somewhat considerable accuracy of 0.8235 and 0.8514 respectively. Moreover, the RB and NN models have accomplished a moderately improved accuracy of 0.8723 and 0.8578 respectively. Furthermore, the MGMM-KSVM technique has resulted in a near optimal performance with an accuracy of 0.9056. However, the proposed ADLVS-FFD technique has attained superior outcomes with a maximum accuracy of 0.9145.

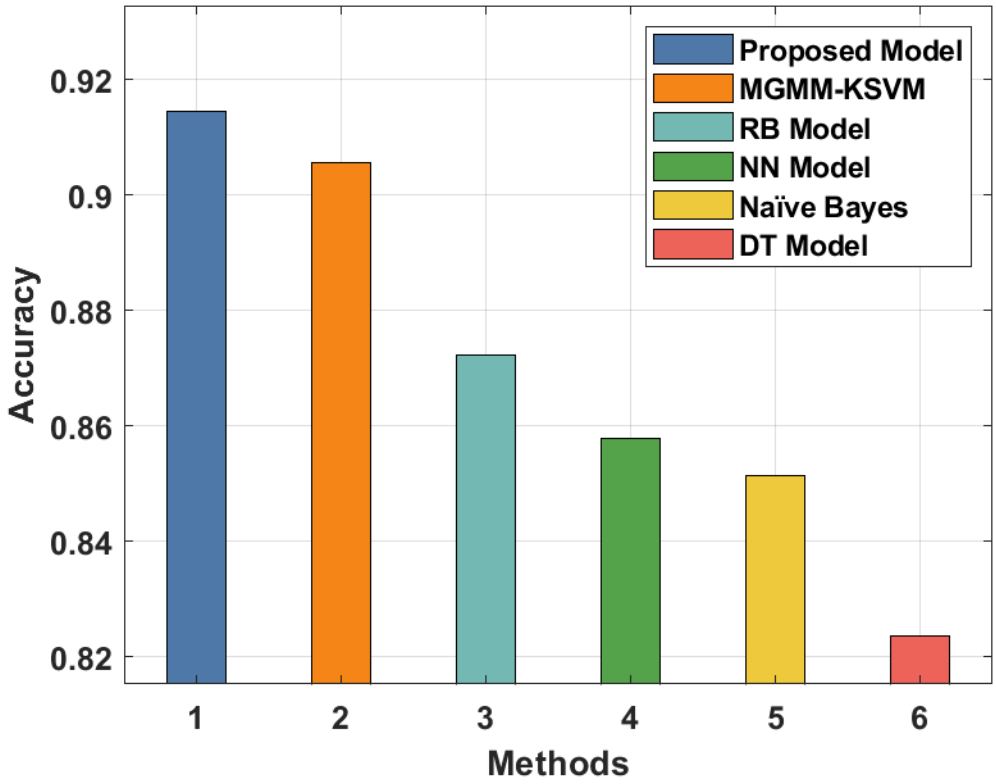

Fig. 4. Accuracy analysis of ADLVS-FFD with existing techniques

4. Conclusion

This paper has developed an optimal ADLVS-FFD technique to summarize videos and detect forest fires. The ADLVS-FFD technique aims to summarize the captured videos and detects the existence of forest fire in it. In addition, the ADLVS-FFD technique involves different subprocesses such as frame splitting, feature extraction, and finally classification. Besides, an MGMM is used for the extraction of key frames and features. Moreover, LSTM model is employed for the detection and classification of input images into normal and forest fire images. In order to ensure the better performance of the ADLVS-FFD technique, a comprehensive experimental validation process takes place on benchmark video dataset. The resultant experimental validation process highlighted the supremacy of the ADLVS-FFD technique over the recent techniques with a maximum accuracy of 0.9145. As a part of future scope, the ADLVS-FFD technique can be implemented in real time environment.

References

[1] He, X., Hua, Y., Song, T., Zhang, Z., Xue, Z., Ma, R., Robertson, N. and Guan, H., 2019, October. Unsupervised video summarization with attentive conditional generative adversarial networks. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2296-2304).

[2] Rani, S. and Kumar, M., 2020. Social media video summarization using multi-Visual features and Kohnen's Self Organizing Map. Information Processing & Management, 57(3), p.102190.

DOI: 10.5281/zenodo.5515606
Received: December 19, 2020 Accepted: August 11, 2021

[3] Khan, G., Jabeen, S., Khan, M.Z., Khan, M.U.G. and Iqbal, R., 2020. Blockchain-enabled deep semantic video-to-video summarization for IoT devices. Computers & Electrical Engineering, 81, p.106524.
[4] Singh, G., Singh, N. and Kumar, K., 2019. PICS: a novel technique for video summarization. In Machine Intelligence and Signal Analysis (pp. 411-421). Springer, Singapore.
[5] Elharrouss, O., Al-Maadeed, N. and Al-Maadeed, S., 2019, June. Video summarization based on motion detection for surveillance systems. In 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC) (pp. 366-371). IEEE.
[6] Chi, R.; Lu, Z.M.; Ji, Q.G. Real-time multi-feature based fire flame detection in video. IET Image Process. 2016, 11, 31–37
[7] Evarts, B. Fire loss in the United States during 2017; National Fire Protection Association, Fire Analysis and Research Division: Quincy, MA, USA, 2018
[8] Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X. and Yao, C., 2018, April. Video summarization via semantic attended networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).
[9] Ji, Z., Ma, Y., Pang, Y. and Li, X., 2019. Query-aware sparse coding for web multi-video summarization. Information Sciences, 478, pp.152-166.
[10] John, A.A., Nair, B.B. and Kumar, P.N., 2017, April. Application of clustering techniques for video summarization–an empirical study. In Computer Science On-line Conference (pp. 494-506). Springer, Cham.
[11] Ji, Z., Xiong, K., Pang, Y. and Li, X., 2019. Video summarization with attention-based encoder–decoder networks. IEEE Transactions on Circuits and Systems for Video Technology, 30(6), pp.1709-1717.
[12] Yuan, L., Tay, F.E.H., Li, P. and Feng, J., 2019. Unsupervised video summarization with cycle-consistent adversarial LSTM networks. IEEE Transactions on Multimedia, 22(10), pp.2711-2722.
[13] Varini, P., Serra, G. and Cucchiara, R., 2017. Personalized egocentric video summarization of cultural tour on user preferences input. IEEE Transactions on Multimedia, 19(12), pp.2832-2845.
[14] Yasmin, G., Chowdhury, S., Nayak, J., Das, P. and Das, A.K., 2021. Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework. Neural Computing and Applications, pp.1-22.
[15] Basavarajaiah, M. and Sharma, P., 2021. GVSUM: generic video summarization using deep visual features. Multimedia Tools and Applications, 80(9), pp.14459-14476.
[16] Davids, D.M. and Christopher, C.S., 2021. An efficient video summarization for surveillance system using normalized k-means and quick sort method. Microprocessors and Microsystems, 83, p.103960.
[17] Messaoud, S., Lourentzou, I., Boughoula, A., Zehni, M., Zhao, Z., Zhai, C. and Schwing, A.G., 2021. DeepQAMVS: Query-Aware Hierarchical Pointer Networks for Multi-Video Summarization. arXiv preprint arXiv:2105.06441.
[18] Parihar, A.S., Pal, J. and Sharma, I., 2021. Multiview video summarization using video partitioning and clustering. Journal of Visual Communication and Image Representation, 74, p.102991.
[19] Zivkovic, Z., 2004, August. Improved adaptive Gaussian mixture model for background subtraction. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 2, pp. 28-31). IEEE.
[20] Ganesh, V. and Kamarasan, M., 2020. Deep learning based long short term memory model for emotions with intensity level sentiment classification for twitter texts. Int. J. Adv. Sci. Technol.
[21] https://github.com/cair/Fire-Detection-Image-Dataset
[22] Pushpa, B. and Kamarasan, M., Video Summarization Based on Gaussian Mixture Model and Kernel Support Vector Machine for Forest Fire Detection, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019

DOI: 10.5281/zenodo.5515606
Received: December 19, 2020 Accepted: August 11, 2021