An Efficient Model for Dynamic Video Summarization based-on Deep Visual Features

Metwally Rashad
Computer Science Department
Computers & Artificial Intelligence,
Benha, Egypt.
metwally.rashad@fci.bu.edu.eg

Gamal El-Nagar
Computer Science Department
Computers & Artificial Intelligence,
Benha, Egypt.
gamal.essam@fci.bu.edu.eg

Ahmed El-Sawy
Computer Science Department
Computers & Artificial Intelligence,
Benha, Egypt.
ahmed.el_sawy @fci.bu.edu.eg

Abstract— **Digital video consists of frames that are presented at variant frame rates such as 15, 24, 30, 60 fps. There is a saying "A picture is worth a thousand words." Pertaining to Digital Video the saying is "a video represents a million of those words strung together". A digital video's core components are frames. Video summarization is the procedure of distilling an original video to build a short raw video that keeps the important stories/content without sacrificing too much information. In this paper, a proposed model is offered for enhancement of dynamic video summarization based on deep learning producing a video skimming summary by selecting keyframes that have importance score met the specification. The task of assigning importance scores to certain frames in a video is critical for summarization, but it is also tough. Proposed model assigns a probability to each video frame, indicating how important it is to be selected, and then selects frames based on the expected values, creating video summaries. Using InceptionV3, ResNet50 pretrained models, it early fuses two feature sets for object-based and place-based in order to predict importance scores. The performance of proposed model has been compared to state-of-the-art models on TVSum dataset. Experimental evaluation show that proposed model performs better in terms of F-score to assess the effectiveness of the proposed model.**

Keywords—Video Summarization, AI, CNN, Image Processing

I. INTRODUCTION

A video is a collection of frames. A digital video's basic parts are called frames. Shots are made from of frames, where the video's content changes dramatically, shots are divided. Scenes are collections of video shots. When the video's background changes or the content changes, the scene changes. Finally, numerous scenes are sequenced to produce the full video.

The development of better video recording and rendering technology has resulted in a significant increase in the volume of video content produced. The video has two issues that make it inconvenient in particular situations. For starters, it necessitates greater storage. Second, in order to determine the content of a video, you must view it thoroughly, which takes too much time. Video summarization developed as a method for facilitating content storage, browsing, and watching without boredom, as well as indexing, retrieval, archiving, and sharing.

The procedure of video summarization starts with cutting video into frames, followed by the choice of the keyframes of the video to form the summary. Using feature extraction techniques, visual features are recovered from the video frames , features to be extracted is chosen based on the video type and summary method. AI approaches such as machine-learning and deep learning are used to organize the frames based on their content. The selection of machine learning approach relies on the labeled data, supervised learning methods are used when the data is labeled, and unsupervised methods are used when labeled data isn't available. The video is then summarised by skimming keyframes. The duplication reduction technique can be applied to improve the summary's quality by deleting redundant frames. Feature extraction is the most important step in the summarising process. There are two strategies to derive these frames' features.By applying image processing algorithms to these frames, or by utilising a machine learning models, like Convolutional Neural Networks (CNNs), that produces deep features after being trained on a large number images. The proposed model will use the 'InceptionV3' and 'ResNet50' pretrained CNN models .

There are four different forms of summarized video. Static video summary A "storyboard" is a sequential grouping of key frames from a video. The storyboard is created by utilizing a specific technique to choose important frames based on a parameter or the user's interest. Dynamic video summary "Video skimming" is the act of constructing a video summary by picking video clips that are of interest to users or snippets that include the video's most important information. Panoramic image summary A "mosaic" is a panoramic image in which the flow of events is represented by a single image or a collection of panoramic images. Summary in "Text" These are paragraph-length summaries that simply provide the textual description of a video sequence. It is generated using Natural Language Processing (NLP) techniques and has no audio or visual description. The model will create a video skimming summary rather than a storyboard summary using proposed model.

II. RELATED WORKS

The authors in [1] suggest a video summarization method based on colour histogram key frame extraction that creates a

978-1-6654-8808-2/22/$31.00 ©2022 IEEE

static video summary. They used colour in key frame extraction since colour is a crucial element for video frames. The initial key frame is manually selected. Because of its simplicity and precision, the colour histogram is a popular way for describing the colour characteristic in a frame. Each frame's colour histogram is calculated. The histogram difference between two successive frames is then determined using the Euclidean distance algorithm. It is determined the distance between two colour histograms. The current frame is chosen as the last key frame if the Euclidean distance is greater than a predefined threshold. The length of key frames can be controlled by changing the threshold value. A higher threshold results in fewer important frames. A lower threshold, on the other hand, will result in more important frames. Instead of utilizing the histogram differences of two frames in each repetition of key frame extraction, the threshold is established using the mean and standard deviation of histogram difference.

In [4] was proposed to estimate probabilities for video frames and make frame selection decisions based on the projected probability distributions; a deep summarization network (DSN). It lays out an end-to-end, reinforcement learning-based methodology for training DSN, in which they create a diversity-representativeness reward function that directly analyses how varied and representative the generated summaries are. It uses CNN to extract visual features from the input video frames in the encoder following by a bidirectional recurrent neural network (BiRNN) with a fully connected (FC) layer with sigmoid function that predicts a probability for each frame, from which a frame-selection action is sampled and a video summary is produced of the selected frames.

For creating static summaries for input videos, [12] developed a technique. Users are only interested in a synopsis if it keeps the important information of the original video. For videos from many genres, the approach can provide reliable and consistent results. The frame set corresponding to the input video is processed to reduce the number of redundant frames. CNN is used to extract high-level feature vectors from the generated set. To represent frames, the layer activations of fully connected layers of pretrained models are used. Using a combination of features extracted using four pre-trained CNN models improves the usefulness of feature vectors. The feature vectors from four models are then concatenated and input into SAE to create a composite representation. The Random Forest classifier is then used to identify keyframes, resulting in a static summary video.

The [5] presented approach for unsupervised video summarising that extracts key-shots from an input video automatically. Based on its empirical findings, there were two important difficulties. The first is poor feature learning caused by flat output significance score distributions for each frame, and the second is training difficulty when dealing with long-length video inputs. It introduces a simple yet effective regularisation loss term called variance loss to minimise the first difficulty. The suggested variance loss enables a network to predict output scores for each frame with large disparity, allowing for effective feature learning and improving model performance dramatically. For the second challenge, it devises a novel two-stream network known as the Chunk and Stride Network "CSNet," which employs both local (chunk) and global (stride) temporal views on video features.

The proposed MSVA model in [22] has multiple sources of visual features where attention is applied to each source in a parallel fashion. The MAVS system is a global attention memory enhanced video summarizer. It combines three different feature types to create a deeper representation of frames in videos; it uses the pretrained 'GoogleNet' model for visual and content-based features extraction, and the pretrained I3D model for motion-related features (Inflated 3D ConvNet). Each type of feature is fed into its own attention mechanism, which is then combined with two linear and single normalization layers to build implicit representations of each type of feature. Finally, the result vector is passed into a sigmoid function, which generates relevance scores for the input frames before generating dynamic video summaries depending on particular parameters.

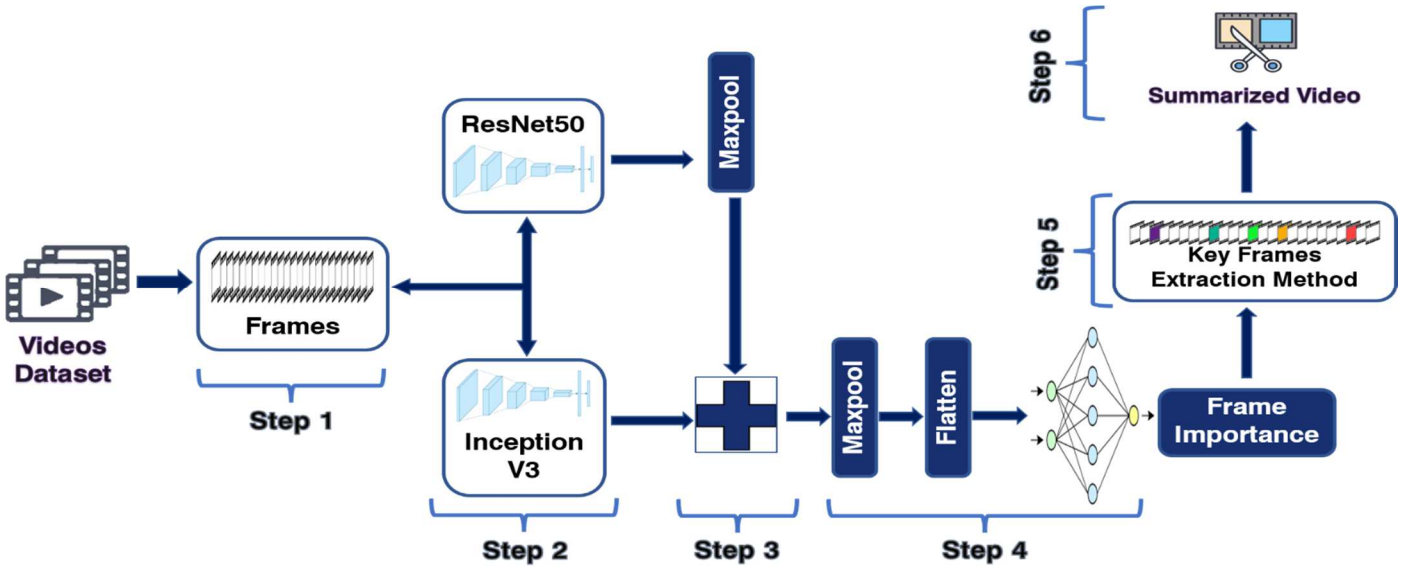

Figure.1 Proposed Model Framework

III. PROPOSED MODEL

In this paper, an efficient model for dynamic video summarization using deep learning is proposed, as shown in Figure.1, it illustrates the steps of proposed model to produce dynamic summarized video. Initially, each video will split into frames (step 1). In step 2, features are extracted from two source of

CNN pre-trained models which are 'ResNet50' pre-trained on Places365 which used in extract visual scene and places features in frames and 'InceptionV3' which is object-based feature extractor. So, let the feature vector from 'ResNet50' be 'V_{f1}' and its shape 'Sh_{f1}' and from 'InceptionV3' be 'V_{f2}' and its shape 'Sh_{f2}'.

In step 3, feature vectors 'V_{f1}', 'V_{f2}' are concatenated but do this, must ensure that the shape of two vectors are same, and to do this, model fed 'V_{f1}' to maxpool layer to reduce dimension and reshape the feature vector from 'ResNet50' 'V_{f1}' to be same as shape of feature vector from 'InceptionV3', it also helps to reduce overfitting. So now '$Sh_{f1} = Sh_{f2}$' and process of early fusion is done producing a new feature vector that contains 'V_{f1}, V_{f2}', let it be 'V_c'. Concatenation of features can help to reduce the number of parameters in model, it can speed up the model training.

In step 4, model will pass final feature vector 'V_c' to maxpool layer to reduce overfitting then to flatten layer. Flattening is the last step performed in a CNN. It involves taking the pooled feature map that is generated in the pooling step and transforming it into a one-dimensional vector. This is done so that you can feed them as inputs to the 'ANN' that has dense layer contains 128 neurons to reduce the dimensions to make the model faster then, the output layer that make prediction of frame's importance score.

In step 5, after proposed model predicts importance score for each frame in a video. Let 'PS_i' be prediction score and 'S_i' be ground truth importance score for frame 'F_i' where '$i = 1 \rightarrow N$ ' where N is total number of frames per Video 'V', each video in dataset has ground truth for its frame's importance score. In proposed model, the extraction of key frames from video done by the following method, let '$T1$' is threshold value and it comes from Equation 1; then, if Equation 2 achieved the frame 'F_i' will be selected as keyframe and added to keyframes vector 'K_f' and its length will be less than N.

$$T1 = \frac{\sum_{i=1}^{N} PS_i}{N} \tag{1}$$

$$\begin{Bmatrix} if\ PS_i > T1 & K_f.add(F_i) \\ Otherwise & discard\ F_i \end{Bmatrix} \tag{2}$$

At the final step (step 6), summarized video will be constructed from selected frame in K_f to produce ' SV ' summarized video with length less than N and equal to length of K_f and its frame rate will be same frame rate of original video 'V'.

IV. EXPERIMENTAL RESULTS

A. *Dataset*

For experimentation, The Title-based Video Summarizing "TVSum" [23] dataset is used to evaluate video summarization techniques. It includes 50 videos ranging in length from two to ten minutes, annotated by 20 people and covering a variety of genres (e.g., news, how-to, documentary, vlog, egotistical).

B. *Evaluation Methods*

The TVSum dataset provides ground truth scores to each annotated frames 'S_i'. The evaluation of proposed model done by measuring the F-score from the set of frames that have been selected by the model. Let 'GSV' ground truth summarized video that will be constructed from keyframes vector 'GK_f' by selecting frame 'F_i' that ground truth importance score 'S_i' greater than threshold value 'T2' as in Equation 3 and Equation 4.

$$T2 = \frac{\sum_{i=1}^{N} S_i}{N} \tag{3}$$

$$\begin{Bmatrix} if\ S_i > T2 & GK_f.add(F_i) \\ Otherwise & discard\ F_i \end{Bmatrix} \tag{4}$$

Now let overlapping frames between 'SV' and 'GSV' '$Overlap_f$' as in eq.4

$$Overlap_f = \{F \rightarrow F \in SV \cap GSV\} \tag{5}$$

The metrics are calculated as given below. Let the summarized video 'SV' consists of 'N_{sv}' frames. Let '$N_{overlap}$' represents the frames in the generated summary that matches the frames in the ground truth summary 'GSV' and let 'N_{GSV}' be the frames in the ground truth summary. Then

$$Precision = \frac{N_{overlap}}{N_{SV}} \tag{6}$$

$$Recall = \frac{N_{overlap}}{N_{GSV}} \tag{7}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

Precision, Recall, and F-measure measure the temporal coherence between summarised video and ground truth summarised video to estimate summary quality. They ignore the more sophisticated human preferences for video shots. To solve this issue, Kendall's '$\boldsymbol{\tau}$' and Spearman's rank-based '$\boldsymbol{\rho}$' metrics are used in this section to give a full evaluation of summary quality. They measure the correlation between the predicted importance curve and ground truth importance curve.

The evaluation of proposed model on TVSum datasets is based on k-fold cross-validation where k = 5. As shown in Table.1

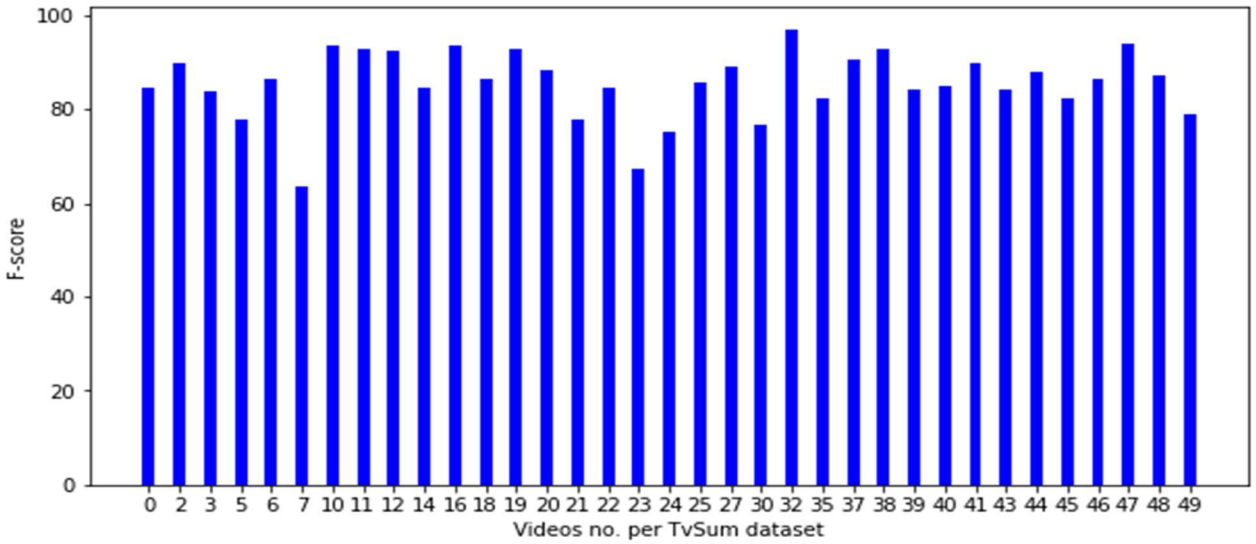

Figure.2 F-scores for videos in TVSum Dataset.

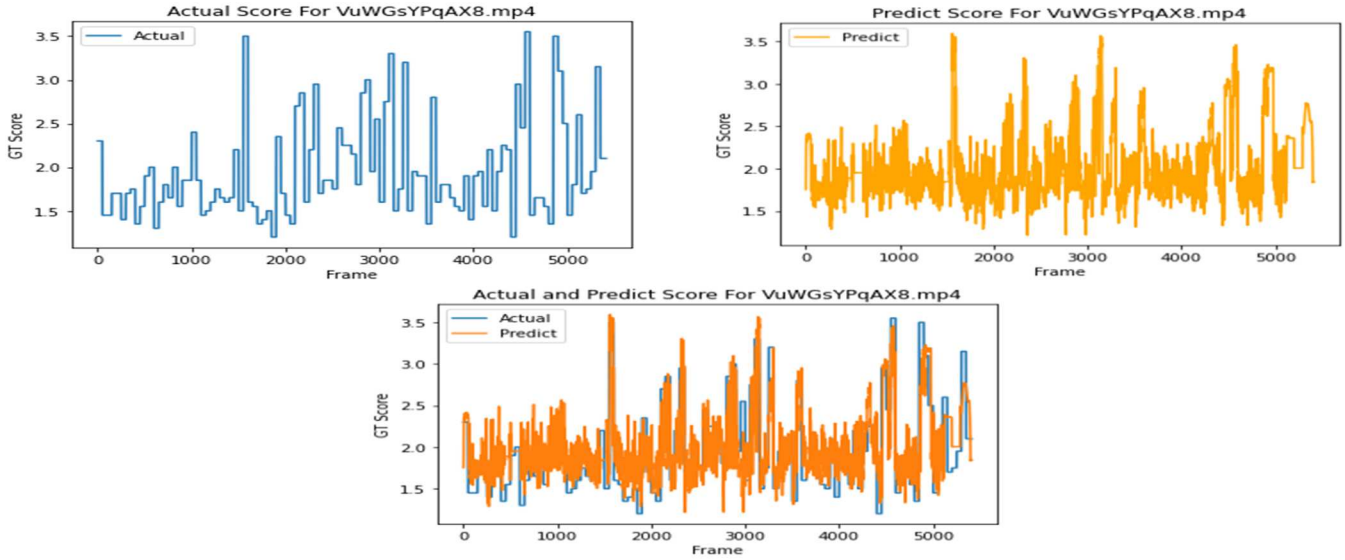

Figure.3 The ground truth curves between actual and prediction scores for video no.30 in TVSum Dataset.

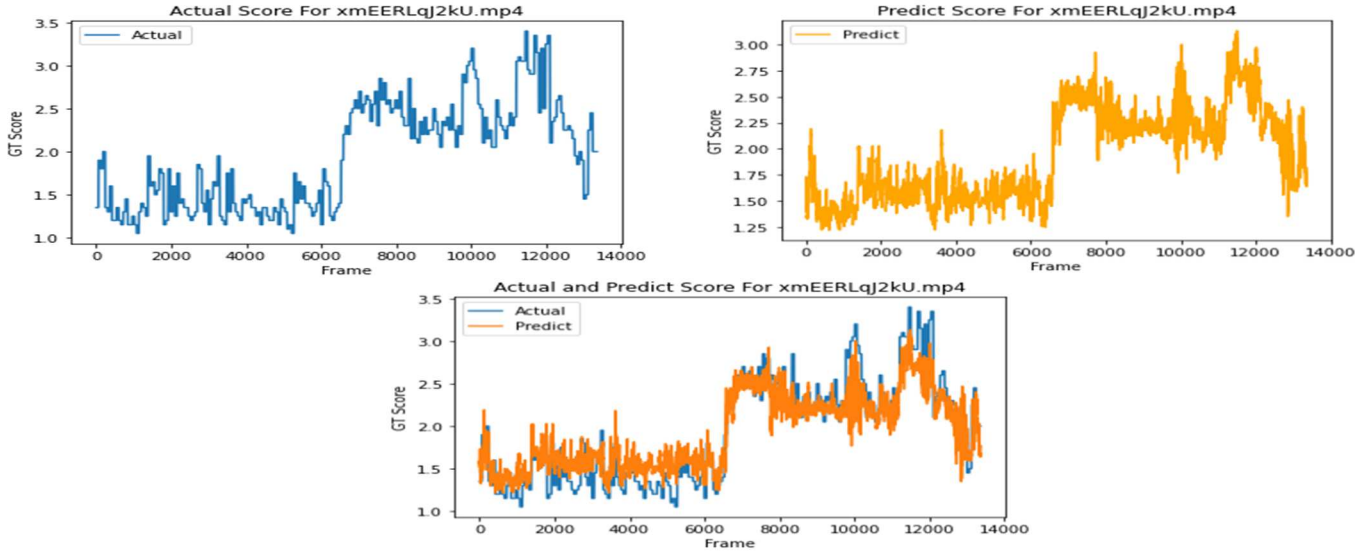

Figure.4 The ground truth curves between actual and prediction scores for video no.32 in TVSum Dataset.

TABLE.1 THE RESULT OF EVALUATION METHODS OF EACH FOLD FOR PROPOSED MODEL.

Fold	TVSum		
	F-score	Spearman's ρ	Kendall's τ
1	80.93	82.80	65.33
2	84.43	86.65	69.80
3	79.44	82.63	65.41
4	83.90	82.34	64.66
5	79.68	77.73	60.64
Avg	81.67	82.43	65.17

C. *Discussion and Results*

The 5-fold cross validation was done by traditional way with shuffle splits which leads to overlapping splits and also there are thirty-five videos are used in testing and as shown in figure.2, it illustrates F-score analysis for these videos in TVSum. The scores are taken when videos were used for testing across 5-folds.

For more visualize results figure.3, figure.4 illustrates the actual score curve and prediction score curve for sample from these videos.

Table.2 shows the results of proposed model compared with other state-of-art that use same dataset with 80% train and 20% test, giving better results than all compared models with F-score 81.67 which means it's about 20% higher.

TABLE.2 EVALUATION OF PROPOSED MODEL COMPARING TO STATE-OF-ART

Models	TVSum
	F-score
$SUM - GAN_{sup}$ [2]	56.3
M-AVS [10]	61.0
VASNet [19]	61.4
MC-VSA [13]	63.7
re-SEQ2SEQ [14]	63.9
MAVS [16]	67.5
MSVA [22]	62.8
The proposed model	**81.67**

V. COCOLUSION

An efficient model that use deep visual features for enhancement of dynamic video summarization is proposed. The proposed model is based on assessing an importance frame using features were extracted from 'InceptionV3' and 'ResNet50' pretrained model following by early fusion process then maxpooling and flatten layer to prepare input to ANN that predict the frame important score. Proposed model is evaluated with 3 metrics and comparing with other models. The results demonstrate great advancement and high efficiency when comparing to other modern models.

VI. REFERENCES

[1] Sarmadi, S. (2017). New Approach in Video Summarization Based on Color Feature. Bulletin de la Société Royale des Sciences de Liège.

[2] Mahasseni, Behrooz & Lam, Michael & Todorovic, Sinisa. (2017). Unsupervised Video Summarization with Adversarial LSTM Networks. 10.1109/CVPR.2017.318.

[3] Asim, Muhammad & Almaadeed, Noor & Al-ma'adeed, Somaya & Bouridane, Ahmed & Beghdadi, Azeddine. (2018). A Key Frame Based Video Summarization using Color Features. 1-6. 10.1109/CVCS.2018.8496473.

[4] Zhou, Kaiyang & Qiao, Yu. (2017). Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward.

[5] Jung, Y., Cho, D., Kim, D., Woo, S. and Kweon, I., 2019. Discriminative Feature Learning for Unsupervised Video Summarization. Proceedings of the AAAI Conference on Artificial Intelligence, 33, pp.8537-8544.

[6] Liang, Buyun & Li, Na & He, Zheng & Wang, Zhongyuan & Fu, Youming & Lu, Tao. (2021). News Video Summarization Combining SURF and Color Histogram Features. Entropy. 23. 982. 10.3390/e23080982..

[7] Vamsi, R. & Subburaman, Dhivya. (2022). A Review on Video Summarization. 10.1007/978-981-16-5652-1_44.

[8] Chakraborty, Saikat. (2019). A Graph-based Ranking Approach to Extract Key-frames for Static Video Summarization.

[9] Elfeki, Mohamed & Borji, Ali. (2019). Video Summarization Via Actionness Ranking. 754-763. 10.1109/WACV.2019.00085.

[10] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li, "Video summarization with attention-based encoderdecoder networks," IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 6, pp. 1709–1717, 2020

[11] Sridevi, M. and Kharde, M., 2020. Video Summarization Using Highlight Detection and Pairwise Deep Ranking Model. Procedia Computer Science, 167, pp.1839-1848.

[12] Nair, M. and Mohan, J., 2020. Static video summarization using multi-CNN with sparse autoencoder and random forest classifier.

[13] Yen-Ting Liu, Yu-Jhe Li, and Yu-Chiang Frank Wang,"Transforming multi-concept attention into video summarization," in ACCV - 15th Asian Conference on Computer Vision. 2020, vol. 12626, pp. 498–513, Springer.nd Video Processing, 15(4), pp.735-742.

[14] Ke Zhang, Kristen Grauman, and Fei Sha, "Retrospective encoders for video summarization," in ECCV - 15th European Conference on Computer Vision. 2018, vol. 11212, pp. 391–408, Springer.

[15] Hussain Kanafani, Junaid Ahmed Ghauri, Sherzod Hakimov, Ralph Ewerth. 2021. Unsupervised Video Summarization via Multi-source Features. In Proceedings of the 2021 International Conference on MultimediaRetrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan. ACM, New York, NY, USA, https://doi.org/10.1145/3460426.3463597

[16] Litong Feng, Ziyin Li, and Zhanghui Kuang et al.,"Extractive video summarizer with memory augmented neural networks," in ACM Multimedia Conference on Multimedia Conference, MM. 2018, pp. 976–983, ACM.

[17] Apostolidis, Evlampios & Adamantidou, Eleni & Metsai, Alexandros & Mezaris, Vasileios & Patras, Ioannis. (2020). AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization. IEEE Transactions on Circuits and Systems for Video Technology. PP. 1-1. 10.1109/TCSVT.2020.3037883.

[18] Li, Ping & Tang, Chao & Xu, Xianghua. (2021). Video summarization with a graph convolutional attention network. Frontiers of Information Technology & Electronic Engineering. 22. 902-913. 10.1631/FITEE.2000429.

[19] Jiri Fajtl, Hajar Sadeghi Sokeh, and Vasileios Argyriou et al., “Summarizing videos with attention,” in ACCV Workshops - 14th Asian Conference on Computer Vision. 2018, vol. 11367, pp. 39–54, Springer.

[20] Zhao, Bin & Li, Haopeng & Lu, Xiaoqiang & Li, Xuelong. (2021). Reconstructive Sequence-Graph Network for Video Summarization.

[21] Gao, Yongbiao & Xu, Ning & Geng, Xin. (2021). Video Summarization via Label Distributions Dual-Reward. 2403-2409. 10.24963/ijcai.2021/331.

[22] Ghauri, Junaid & Hakimov, Sherzod & Ewerth, Ralph. (2021). Supervised Video Summarization Via Multiple Feature Sets with Parallel Attention. 1-6s. 10.1109/ICME51207.2021.9428318.

[23] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes, “TVSum: Summarizing web videos using titles,” in IEEE Conference on Computer Vision and Pattern Recognition, CVPR. 2015, pp. 5179 5187,IEEE Computer Society.