Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 10, No. 1, October : 2023

OBSERVATIONAL EXAMINATION OF COMPUTATIONAL EXECUTION FOR RECORDINGS SUMMAZATION MODELS

Miss Hemlata More RCET Bhilai, India hemamore41@gmail.com
Dr. Pankaj Kumar Mishra RCET Bhilai, India bpmishra1974@yahoo.co
Mrs Lakhwinder Kaur RCET Bhilai, India lakhwinder20063@yahoo.com

Abstract:
Summarization of large length videos is a multidomain task which includes video pre-processing, feature extraction, feature analysis, and post processing. Each of these tasks requires large computational delays, which limits their performance for long length video sequences. Moreover, after summarization all key events present in input video should be reflected at output, while keeping output video size as low as possible. High speed algorithms produce large sized summarized videos, which limits their usability for event-based summarization applications. In order to remove these drawbacks, this text proposes a long-short-term-memory (LSTM) based event detection model, which estimates start and end timestamps for events. Each of these event sequences is then given to a variance-based model for evaluation of intra-event keyframes. In order to perform this task, a convolutional neural network (CNN) is deployed for estimation of events, followed by a variance thresholding model for selection of keyframes from each event set. Due to this approach, the accuracy of video summarization is improved by 5%, while the compression ratio is improved by 8% when compared with state-of-the-art models. These metric improvements allow the proposed system to be used for a wide variety of video summarization applications including closed circuit television (CCTV), military surveillance, sports, and other event-based videos. The proposed model is evaluated on standard TRECVID dataset, and possesses an accuracy of 98% on different video types.

Keywords: Video, summarization, compression, accuracy, CNN, LSTM

Introduction

Video summarization for large-sized sequences requires efficient design of various image processing models. These models include, but are not limited to, shot boundary detection, clustering of different shot boundaries, event classification, feature estimation from these shot boundaries & events, and final summarization. This process converts large-sized videos into smaller chunks, via event timestamp estimation. Each of these smaller chunks are further analyzed using various feature extraction and classification techniques for effective summarization. In order to perform this task, a wide variety of event-based keyframe extraction (KFE) or video summarization models are designed by researchers. Such an event-based summarization model can be observed from figure 1, wherein motion estimation and object detection are used in order to evaluate position of keyframes in the input video. These keyframes are given to a tracking module, which estimates presence of these frames in the entire video. Finally, an event detection model is activated for identification and grouping of keyframes w.r.t. their event occurrence. This process is combined with coverage summarization, wherein event meta data is exploited in order to improve video summarization efficiency.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 10, No. 1, October : 2023

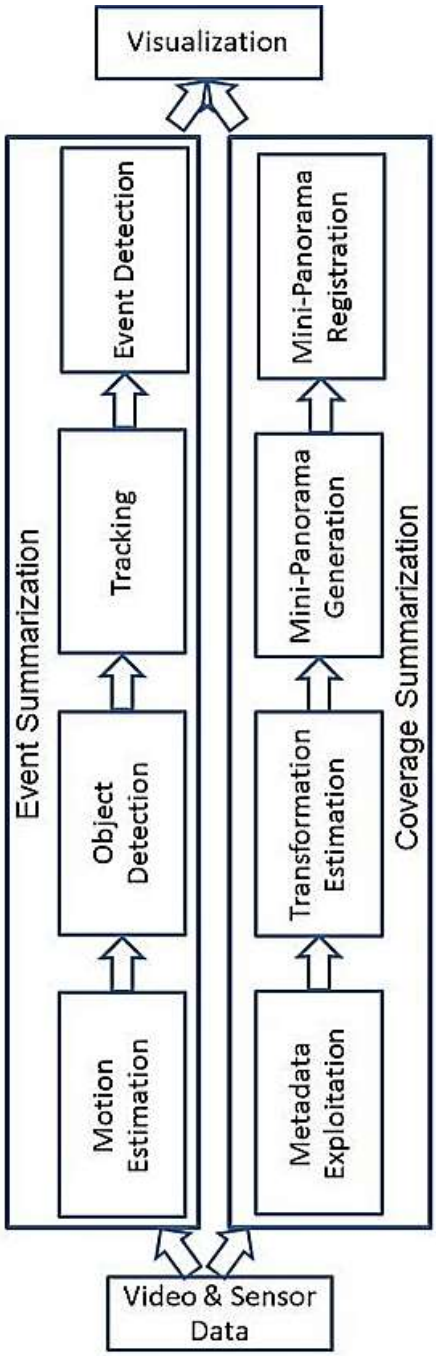

Figure 1. Video summarization using event-based classification

These models are highly effective due to their event-based frame categorization capabilities. The next section describes similar models, and estimates their nuances, advantages and limitations, which assists in designing of the underlying video summarization model. This is followed by design of the proposed model and its performance evaluation when compared with standard summarization models. Finally, this text concludes with some interesting observations about the proposed model, and recommends ways to improve the same.

Literature review

Video outline techniques ordinarily recognize frames that have higher development esteems, and thusly moving article recognition can be a component vector for this reason. The work in [1] utilizes

Moving Object Detection and Image Similarity with the end goal of KFE. It utilizes ViBe calculation and breakers between frame distinction strategy by isolating the first video into a few portions that contain the moving item. For highlight extraction Speed Up Robust Features or SURF is utilized. At long last, a versatile choice edge is utilized for discovering key frames from the information set of frames. They likewise utilize the idea of Peak Signal to Noise Ratio (PSNR) to discover frame comparability. The framework first takes the video frames as info and discovers moving articles from these frames. These article frames are given to PSNR calculation, and a worldwide PSNR likeness highlight is assessed. Neighborhood likeness is discovered utilizing SIFT procedure, and afterward a weighted combination highlight set is assessed to discover the upsides of comparability between the frames. This closeness is contrasted utilizing versatile edge determination with at long last assess the key frames. Utilizing the proposed technique an exactness of 99% can be accomplished. Yet, this is tried on a restricted arrangement of video groupings, it is suggested that the perusers should perform further ingenuity prior to utilizing this strategy for their own exploration. Another 2 level KFE strategy is proposed in [2], [3] and [4], wherein scientists have utilized the idea of CBD and ECR for KFE. Results grandstand that the proposed technique has great KFE execution.

KFE in 3D videos has been a subject of study for over 10 years now. Strategies like shot limit identification (SBD), uniform examining, position testing, bunching (k-Means and FCM), Curve improvement, Minimum connection, Minimum recreation mistake, Matrix factorization are concentrated in [5] and [6]. From the assessment done in [7] and [8], it is tracked down that the SBD strategies beat different techniques for 3D KFE. SBD strategies additionally utilize Uniform Local Binary Pattern (ULBP) to discover key frames. The work in [9] uses Uniform Local Binary Patterns between various frames of a video, and afterward an edge is applied to these frames to discover the vital frames from the given video succession. The proposed strategy outflanks SIFT, Unsupervised and different strategies as far as exactness, review, F-measure, figure of legitimacy and precision. Changing over video arrangements into various shading spaces like HSV, LAB, and others can fundamentally further develop the KFE execution. The work done in [10] changes over the frame groupings into a rundown space, which is a bunched space shaped after use of profound highlights on the information frames. This synopsis space is a demonstrative of the transient contrasts between the frames. The chose weighting strategy is utilized to track down the most pertinent frames utilizing a frame thresholding method. The proposed approach beats bunching based methodology, word reference-based methodology and article put together methodologies when applied with respect to open video dataset and YouTube video datasets. Another audit of KFE for 3D videos is depicted in [11], which demonstrates that the presentation of SBD strategies is superior to any of different techniques. SBD and different methodologies generally use include planning with the end goal of KFE, these highlights can be histogram, shading guides, or complex SURF highlights. The SURF highlights beat other component vectors as far as productivity of key frame extraction. The work in [12] uses the idea of SURF for highlight extraction, and approves that SURF beats other component extraction strategies for KFE. It is suggested that SURF highlights be joined with profound CNNs to assess their presentation as an incorporated KFE calculation. Also, procedures like hash guide can be used with the end goal of KFE as depicted in [13]. In [13], scientists have consolidated hash maps with a limit based classifier to discover the critical frames from given video arrangement. There is no remark on the factual execution of this calculation, and it is suggested that scientists should assess the presentation of this novel strategy prior to utilizing something similar. Another versatile bunching based technique is depicted in [14] that demonstrates that the presentation of HSV with versatile grouping is prevalent than others for KFE.

Security in KFE has consistently been an optional thought boundary. Be that as it may, for exceptionally secure applications like military and medical services, it is needed to have a specific degree of safety during KFE measure. The work in [15] proposes the idea of video parceling for getting

KFE for industry standard MPEG video groupings. It again utilizes SBD for key frame extraction, yet adds the idea of hashing and crypto-space activities for getting the KFE interaction. This strategy can decrease the video grouping length by over 90%. News videos are normally long successions that recurrent data in a circle. The work done in [16] applies pixel contrast between 2 sequential frames to perform KFE. This pixel differencing procedure is joined with text division, and on the off chance that the content qualities change, the frames are set apart as introductory key frames. These underlying key-frames are then given to a k-Means calculation to discover the frame contrast lastly get the critical frames from the video sets. A comparable work is acted in [17], that uses visual consideration model and movement energy vectors to discover key frames. The visual consideration model is conceived utilizing worldwide closeness include like shading histogram, object shape, optical stream and different techniques. Then, at that point PSNR esteems are assessed to discover the frame similitude. Disparate frames are set apart as starting key frames, and afterward are utilized for the second phase of handling. In the subsequent stage, neighborhood similitude highlights are assessed utilizing Harris Corner Detection and optical stream. In view of these highlights, and a limit indicator the ideal key frames are separated. The outcomes grandstand that the proposed calculation can discover key frames with over 90% effectiveness. Another shading histogram-based technique is proposed in [18], wherein the HSV model is joined with shading space quantization, and a shading histogram is assessed. Then, at that point utilizing Euclidean distance between the histograms of progressive frames, the key frames are removed. The outcomes are assessed on facial videos, and it is tracked down that the pace of facial acknowledgment improves by over 10% when contrasted with different strategies. A word reference-based technique is depicted in [19], wherein Latent Semantic Analysis (LSA) is applied alongside significant level indicators to assess key frames. Here, subtractive grouping is utilized to discover the most variation frames from the given arrangement of info frames.

Proposed event-based model for video summarization using LSTM

The proposed event-based model for long video summarization uses a combination of LSTM based CNN with feature variance method for estimation of keyframes. The model works is 2-phases, wherein in the first phase various events are recognized from the video using LSTM based CNN model as observed from figure 2; while in the second phase these events are individually summarized using variance-based threshold engine.

The LSTM CNN model for event estimation starts by taking the input training video, and converting it into different event parts. Each of these event parts are given to individual convolutional layers for event-based feature extraction. This enables the LSTM CNN model to analyze different patterns, and use these patterns for estimation of events in new videos. The extracted features are then given to individual max pooling layers, wherein feature reduction and selection processes are performed. This layer reduces the number of extracted features to nearly half, thereby removing feature redundancy, improving accuracy of classification and reducing the delay needed for training and classification.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 10, No. 1, October : 2023

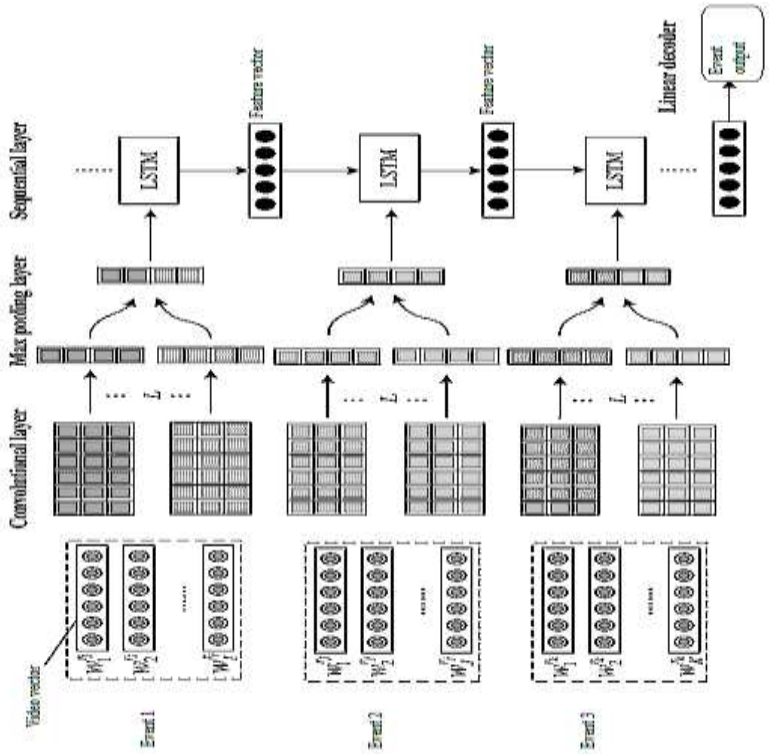

Figure 2. Proposed LSTM CNN Model for event estimation from input videos

Results of the max pooling layer are given to different LSTM model blocks, wherein event specific models are generated. Each generated LSTM model block represents an individual event which can be recognized by the system. Finally, a linear decoder module is used for finding out type of event in the input video.

Results of LSTM CNN model are given to a feature variance model. The feature variance model initially evaluates extended histogram and edge map of each event video; which allows it to estimate colour and shape features. These features are then given to a feature variance model for estimation of the most variant features. Frames that have most variant features are produced at the output, while minimal variance frames are removed from the video. In order to find out extended histogram and edge maps, the following process is followed,

- Non-identical pixel colour levels are extracted from input image, let 'N' be the number of colour levels extracted
- Colour histogram ($Hist_{color}$) is estimated by counting the number of pixels of that particular colour as present in the input image. This process is showcased using equation 1,

$$Hist_{color_j} = \prod_{i=1}^{N} I_{i_j} \dots (1)$$

- After this, let unique pixels pairs for this event video be I1:I2, then for each pixel pair use equation 2 for estimation of edge map (EMap) for the given event,

$$EMap_i = \prod_{j=1}^{N} Sobel[I_{I1:I2}] \dots (2)$$

where, 'j' is number of non-identical pixel levels in input image.

Both extended histogram and edge maps are combined together to form the final feature vector. This feature vector is formed for each frame in each event, and then an intra-event variance is estimated using equation 3. Frames which have a variance less than this intra-event value are removed from

the event sequence, while other frames are marked as summary frames and are presented at the output.

$$V_{avg} = \sqrt{\frac{\sum_{a=1}^{m}(x_a - \frac{\sum_{i=1}^{m}\sqrt{\frac{\sum_{j=1}^{n}(x_j - \frac{\sum_{k=1}^{n} x_k}{n})^2}{n-1}}}{m})^2}{m-1}} \quad \ldots (3)$$

Where, 'm' is the number of frames in the current event class, 'n' is number of frames in the other event class, and 'x' is combination of extended histogram map and edge map feature vectors. Based on this evaluation, summaries are generated for different videos from the TRECVID dataset. Evaluation of this model on different videos, and parametric evaluation can be observed from the next section, wherein estimation of compression rate and accuracy of summarization are tabulated, analyzed and compared with different standard models.

Result and comparative analysis

Compression rate (CR) and accuracy (A) performance of the proposed LSTM CNN based event classification model with feature variance estimation is estimated for different algorithms. These values are tabulated in table 1, and 2 wherein it can be observed that the proposed model outperforms other models in terms of compression rate, and accuracy values. These values were obtained by evaluating the system on TRECVID dataset, which consists of over 5000 videos in different categories.

Number of videos	Avg. A [6]	Avg. A [15]	Avg. A [Proposed]
10	0.71	0.80	0.80
20	0.76	0.80	0.85
30	0.60	0.85	0.86
40	0.63	0.84	0.85
50	0.71	0.76	0.88
60	0.75	0.78	0.87
70	0.78	0.77	0.88
80	0.71	0.77	0.88
90	0.74	0.78	0.87
100	0.76	0.79	0.87
110	0.78	0.79	0.88
120	0.79	0.79	0.88
130	0.80	0.80	0.91
140	0.82	0.81	0.89
150	0.80	0.82	0.91
160	0.81	0.83	0.91
170	0.82	0.83	0.92
180	0.83	0.84	0.94
190	0.84	0.85	0.94
200	0.84	0.86	0.97
210	0.85	0.86	0.98

Table 1. Accuracy of different video summarization models

From the table we can observe that the proposed system outperforms other state-of-the art methods by almost 5%. These tests were conducted on a large set of videos taken from the TRECVID dataset, and each of the given algorithms were evaluated on the same dataset. The training set consists of 500 videos of different length, sizes and content, while the testing set consists of 210 videos varying in terms of

length, sizes and content. It is found that the proposed algorithm performed very effectively in terms of accuracy of summarization. This can also be observed from figure 3, wherein accuracy values are visualized.

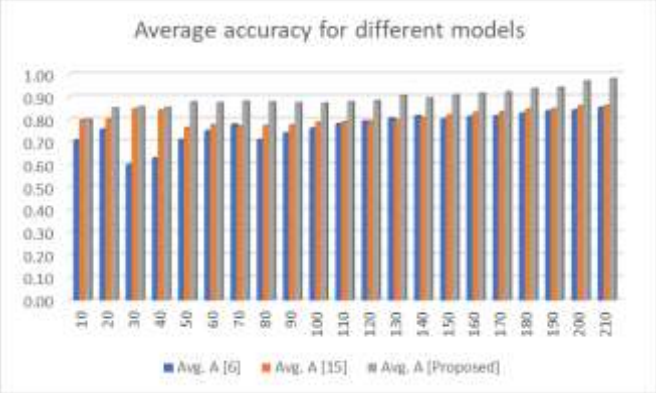

Figure 3. Average accuracy of different models

Similar analysis is done in terms of compression ratio, and can be observed from table 2 as follows,

Number of videos	Avg. CR [6]	Avg. CR [15]	Avg. CR [Proposed]
10	0.40	0.45	0.46
20	0.43	0.46	0.50
30	0.35	0.49	0.51
40	0.37	0.49	0.51
50	0.42	0.45	0.53
60	0.44	0.46	0.54
70	0.46	0.46	0.55
80	0.43	0.46	0.55
90	0.45	0.47	0.56
100	0.46	0.48	0.56
110	0.48	0.48	0.57
120	0.49	0.49	0.58
130	0.50	0.49	0.60
140	0.51	0.50	0.60
150	0.50	0.51	0.62
160	0.51	0.52	0.63
170	0.52	0.53	0.64
180	0.53	0.54	0.65
190	0.54	0.54	0.66
200	0.54	0.55	0.68
210	0.55	0.56	0.69

Table 2. Average compression ratio of different video summarization models

From the delay values it is evident that the proposed event classification-based LSTM CNN model with variance-based summarization outperforms other models, and achieves 8% improvement in terms of compression ratio. This can also be observed from figure 4, wherein these values are visualized. From these results it can be observed that the proposed model highly applicable for real time video summarization applications. As the compression ratio is very high, the proposed model can be used for summarization of large sized video sequences with high accuracy, thereby making it useful for movie summarization, documentary summarization, etc.; which further expands application areas for the proposed combination of LSTM with CNN for efficient summarization of videos.

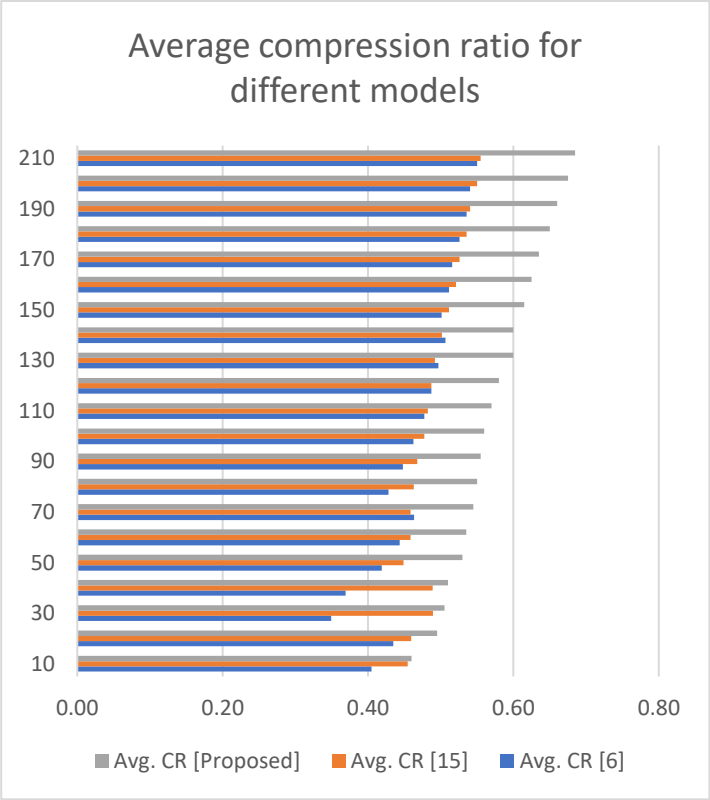

Figure 4. Average compression ratio for different models

Conclusion & Future work

From the results it is evident that the proposed model is better in terms of accuracy, and compression ratio when compared with existing models. These parameters are taken for more than 500 video sequences, and the final test was performed. The proposed LSTM CNN model performs high accuracy event classification, which is followed by feature variance based keyframe selection. Due to this combination of event classification and feature variance selection, the resultant algorithm is superior than other state of the art methods. Due to the use of CNN with LSTM, the delay of initial training is low, thus it further improves algorithmic computational speed. In future, researchers can work on minimizing the delay even further and improving compression ratio for extra-large sized video sequences.

References

[1] Pan, Gang & Zheng, Yaoxian & Zhang, Rufei & Han, Zhenjun & Sun, Di & Qu, Xingming. (2019). A bottom-up summarization algorithm for videos in the wild. EURASIP Journal on Advances in Signal Processing. 2019. 10.1186/s13634-019-0611-y

[2] Apostolidis, Evlampios & Adamantidou, Eleni & Metsai, Alexandros & Mezaris, Vasileios & Patras, Ioannis. (2021). Video Summarization Using Deep Neural Networks: A Survey.

[3] Atencio, Pedro & Sanchez, German & Branch, John & Delrieux, Claudio. (2019). Video Summarization by Deep Visual and Categorical Diversity. IET Computer Vision. 13. 10.1049/iet-cvi.2018.5436.

[4] Z. Ji, Y. Zhao, Y. Pang, X. Li and J. Han, "Deep Attentive Video Summarization With Distribution Consistency Learning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 4, pp. 1765-1775, April 2021, doi: 10.1109/TNNLS.2020.2991083.

[5] L. Yuan, F. E. H. Tay, P. Li and J. Feng, "Unsupervised Video Summarization With Cycle-Consistent Adversarial LSTM Networks," in IEEE Transactions on Multimedia, vol. 22, no. 10, pp. 2711-2722, Oct. 2020, doi: 10.1109/TMM.2019.2959451.

[6] Y. Wang, Y. Dong, S. Guo, Y. Yang and X. Liao, "Latency-Aware Adaptive Video Summarization for Mobile Edge Clouds," in IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1193-1207, May 2020, doi: 10.1109/TMM.2019.2939753.

[7] H. Raksha, G. Namitha and N. Sejal, "Action based Video Summarization," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 457-462, doi: 10.1109/TENCON.2019.8929597.

[8] S. Xiao, Z. Zhao, Z. Zhang, Z. Guan and D. Cai, "Query-Biased Self-Attentive Network for Query-Focused Video Summarization," in IEEE Transactions on Image Processing, vol. 29, pp. 5889-5899, 2020, doi: 10.1109/TIP.2020.2985868.

[9] (2008) Video Segmentation and Keyframe Extraction. In: Machine Learning for Audio, Image and Video Analysis. Advanced Information and Knowledge Processing. Springer, London. https://doi.org/10.1007/978-1-84800-007-0_14

[10] M. Asim, N. Almaadeed, S. Al-maadeed, A. Bouridane and A. Beghdadi, "A Key Frame Based Video Summarization using Color Features," 2018 Colour and Visual Computing Symposium (CVCS), 2018, pp. 1-6, doi: 10.1109/CVCS.2018.8496473.

[11] V. Rajpoot and S. Girase, "A Study on Application Scenario of Video Summarization," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 936-943, doi: 10.1109/ICECA.2018.8474699.

[12] Lopez-Alanis, A., Lizarraga-Morales, R.A., Contreras-Cruz, M.A. *et al.* Rule-based aggregation driven by similar images for visual saliency detection. *Appl Intell* **50,** 1745–1762 (2020). https://doi.org/10.1007/s10489-019-01582-6

[13] Gianluigi, C., Raimondo, S. An innovative algorithm for key frame extraction in video summarization. *J Real-Time Image Proc* **1,** 69–88 (2006). https://doi.org/10.1007/s11554-006-0001-1

[14] A. Bora and S. Sharma, "A Review on Video Summarization Approcahes : Recent Advances and Directions," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 601-606, doi: 10.1109/ICACCCN.2018.8748574.

[15] Z. Zhang, D. Xu, W. Ouyang and C. Tan, "Show, Tell and Summarize: Dense Video Captioning Using Visual Cue Aided Sentence Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 9, pp. 3130-3139, Sept. 2020, doi: 10.1109/TCSVT.2019.2936526.

[16] Y. Yuan, H. Li and Q. Wang, "Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network," in IEEE Access, vol. 7, pp. 64676-64685, 2019, doi: 10.1109/ACCESS.2019.2916989.

[17] X. Ai, Y. Song and Z. Li, "Unsupervised Video Summarization Based on Consistent Clip Generation," 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), 2018, pp. 1-7, doi: 10.1109/BigMM.2018.8499188.

[18] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik and V. H. C. de Albuquerque, "Cloud-Assisted Multiview Video Summarization Using CNN and Bidirectional LSTM," in IEEE Transactions on Industrial Informatics, vol. 16, no. 1, pp. 77-86, Jan. 2020, doi: 10.1109/TII.2019.2929228.

[19] Wang, X., Nie, X., Liu, X. *et al.* Modality correlation-based video summarization. *Multimed Tools Appl* **79,** 33875–33890 (2020). https://doi.org/10.1007/s11042-020-08690-3