ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1011022021.pdf
https://doi.org/10.30534/ijatcse/2021/1011022021

A Survey on various Video Summarization Techniques

Tejeswinee K[1], Bharanidharan M[2], Dinesh Kumar T[3], Arjun Prakash T[4]
[1]Rajalakshmi Engineering College, Anna University, India, tejeswinee.k@rajalakshmi.edu.in
[2]Rajalakshmi Engineering College, Anna University,India, bharanidharanmahesh@gmail.com
[3]Rajalakshmi Engineering College, Anna University, India, dineshdkda31@gmail.com
[4]Rajalakshmi Engineering College, Anna University, India, y2tarjun@gmail.com

ABSTRACT

In this fast-moving world people don't have the time to watch lengthy videos, so it would be convenient for them if they could access short summaries of these videos and be able to acquire more information by watching summarized videos. A study of the different methods used for video summarization over the years i.e., extracting important segments from a video to produce short concise summaries that are representative of the original video is presented.

Key words : Video Summarization, KeyFrame Extraction, Semantic Features.

1. INTRODUCTION

In recent years, video summarization has become a very active research field due to the big amount of video content being generated. The number of videos on the internet and social media are growing at an enormous rate everyday. On average 500 hours of video are uploaded on YouTube every minute. In this era of video boom, it is difficult for individuals to experience most of the videos they like, due to the time it takes to watch them. This gives rise to the need for video summarization. The generation of a compact, comprehensive, and automated summary of video can facilitate an effective way for users to watch their favourite videos.

Video Summarization has applications in many fields. For example, the large amount of video data that is being produced everyday by camera-based systems, such as surveillance, medical and telecommunication systems, makes the job difficult for people who just want to focus on the important parts of the video which are limited. Video Summarization can also be used in/for the sports industry, as the broadcast time is increasing every year. Also, in this fast-moving world people don't have the time to watch lengthy videos, so it would be convenient for them if they could access short summaries of these videos and be able to acquire more information by watching summarized videos.

2. METHODS

Video summarization is the process of extracting important segments from a video to produce short concise summaries that are representative of the original video. Video summarization can be context based and non-context based - context based video summarization techniques understand the needs/requirements of the users and also considers domain specific relationships among video shots, while non-context based summarization techniques focus just on the extraction of keyframes from the original video. Some of the common methods used are:

2.1 Feature Based Summarization

Videos contain many audio and visual features. The most common audio classes in videos are speech, silence, music and the combination of the three, while Color histogram, Edge histogram and Texture similarities are some of the visual features that can be used to extract features. Summarization is done based on the relevant features as per the user's needs. There are several techniques to extract summaries based on an object, event, color, and motion, etc..

2.2 Clustering

There are a number of clustering techniques like similar activities clustering, k-means clustering and spectral clustering. In similar activities clustering, common activities are taken from various frames from the video and features are extracted from them and a highlight video is generated. In k-means clustering histograms are generated for segments of a video using the k-means algorithm and the clusters are grouped together to form a highlighted video. In spectral clustering, the dimensionality of the spectrum is reduced before clustering is performed on the frames.

2.3 Spatio-temporal

This method takes into consideration both space and time. Based on this, nodes are selected from the videos and marked as critical points. Then the objects are detected from the video which are grouped together using a sliding window technique. The object detection is done using clustering algorithm, optical flow, and background subtraction. This technique is useful when the camera is fixed in a position and not in motion..

2.4 Sparse-graph/Dictionary

In this method, data logs are generated and any changes in time are monitored continuously for the videos. The videos can be represented as words in the dictionary. Clustering method and shot boundary technique are used for the words generation. Sparse representation based. In sparse video summarization, frames are divided into patches with the help of features extracted using the convolutional sparse representation and the unique frames are obtained using simultaneous block version of block-based Orthogonal Matching Pursuit (SBOMP) algorithm.

3. LITERATURE SURVEY

Daniel DeMenthon(1998) [1] proposed a system where the videos can be represented in the form of a curve with many dimensional features. These curves are reduced with the help of a binary curve splitting algorithm to a smaller number of points and they are joined to form a polygon that can be represented as a tree structure from which keyframes are extracted. They also developed a video player that allows the user to adjust the level of summarization.This was one of the earliest works in the field of video summarization.

Anastasios D. Doulamis et. al. (2000) [2] proposed a fuzzy representation(mathematical means of representing vagueness and imprecise information) of visual content. In this method, the author uses fuzzy representation of the visual data to perform both summarization and content based indexing and retrieval. A colour/motion segmentation algorithm is used for visual data description and M-RSST(Multiresolution recursive shortest spanning tree) is used for both color and motion segmentation. Then, the segments are classified into classes - to avoid the possibility of classifying two similar segments to different classes a degree of membership is allocated to each class which results in fuzzy classification. Then a fuzzy multidimensional histogram is created. Then, key frames are extracted after removing visually similar frames. For content based retrieval, images are given as input which is then searched in the video database(which contains the keyframes from which M images are retrieved. This approach outperformed the other methods in both accuracy and computational efficiency at the time of publishing.

Yoshimasa Takahashi et. al. (2004) [3] proposed two methods of generating a summary of sports videos. One is to create a concise video clip by temporally compressing the amount of the video data. The other is to provide a video poster by spatially presenting the image keyframes which represent the whole video content. With the help of metadata they summarize the videos in both methods. They got better results than the videos manually summarized by users.

Jian-quan Ouyang et. al. (2005) [4] proposed a system where video abstraction was done by extracting the replays that were available in the video in MPEG format. They use color and camera information to retrieve the keyframes. They calculate the euclidean distance between the frame before the replay scene and the replay frame beginning and if it satisfies the conditions required the frame is selected for summarization. They provided three summarization techniques: One was to include all the replay and live shot scenes. Two was to include all the replay scenes and Three was to include all the live scenes.

G. Evangelopoulos et. al. (2008) [5] proposed a system where both audio and video features were required to produce the summarized video based on saliency that was measured for the audiovisual streams. Audio saliency is obtained from the signal made from any source which is generalized to AM-FM model. Video saliency is obtained by representing the videos in three dimensions from which features are extracted using spatio temporal framework. Both these saliency's are merged in a single attention curve from which the salient events are extracted. This gives an accuracy of 80%.

Robert Laganiere et. al. (2008) [6] proposed an approach composed of five steps. First, image features are collected using the spatio-temporal Hessian matrix. Then, these features are processed to retrieve the candidate video segments for the summary (denoted clips). Further on, two specific steps are designed to first detect the redundant clips, and second to eliminate the clapperboard images. The final step consists in the construction of the final summary which is performed by retaining the clips showing the highest level of activity. This method ranks 24th for redundant frame inclusion and ranks 39 for inclusion criteria.

Jitao Sang et. al. (2010) [7] proposed a system where they segmented the video scenes. Then they aligned the segments with the script of the videos. Then they analyse the substory with character histogram. Content attention analysis is done based on the characters in the video and the number of times they occur in the video. Using the substory discovery and content attention value, they calculate movie attraction scores at both shot and scene levels and adopt this as criterion to generate movie summary. This system is better than the audiovisual feature based summarisation

Sandra Eliza Fontes de Avila et. al. (2011) [8] proposed VSUMM, a method for the generation of static video summaries. The method works on color feature extraction from video frames and k-means clustering algorithm. Additionally, they have also developed a method for evaluating the generated videos. This evaluation method gives us an idea on what basis the methods are compared and evaluated, analyzes the generated video and makes objective comparison between the different methods. This method is compared with user generated highlights and gives a confidence of 98%.

Irfan Mehmood et. al. (2013) [9] proposed a system where they extract the keyframes based on the visual saliency that helps to extract keyframes based on the semantically relevant frames. This method takes into account both low level features and high level features. Visual saliency consists of static saliency and dynamic saliency. The static saliency is derived from color opponent component space using center surround measure, while the dynamic saliency is determined using motion intensity and its phase coherence. The two saliency curves are combined to get a single curve and the peak of the curve was used to extract the keyframes. (Output is given as rating from users). They compared their method with non-visual methods and got a enjoyability and informative score of 89.8 and 92.12.

Karim M. Mahmoud et. al. (2013) [10] proposed a VSCAN method that generates static video summaries. This method is a successor of the DBSCAN clustering algorithm where summarization was based on color and textures features. The VSCAN method produces video summaries with better quality than other methods. This method has a Mean F-Measure of 0.77 which is better than VSUMM, OV, DT, STIMO and BD-COLOR. Further they can add other features like edge and motion descriptors in the future versions of this model. The output of this method can be used as input for video skimming models.

Bin Zhao et. al. (2014) [11] proposed a system where the method learns a dictionary from a given video using group sparse coding. A summary video is then generated by combining segments that cannot be sparsely reconstructed using the learned dictionary. This method can generate the frames needed for the summary on the fly without going through all the frames once before selecting them. Therefore the time required by their method to generate a summary of a video is the same as the length of the video. Here they use an algorithm called LiveLight. LiveLight shows better performance than the state of the art DSVS algorithm by 8%.

A. Ravento et. al. (2014) [12] proposed an audio-visual descriptor based method for Automatic Soccer Highlights summarization. The video is segmented and the duido is matched with the segments and a set of low and mid level descriptors are computed. The final summary is generated by selecting those shots with highest interest according to the specifications of the user and the results of relevance measures. The proposed system is able to detect satisfactorily over the 70% of the total amount of goals of the five soccer matches analyzed.

Michael Gygli et. al. (2014) [13] proposed a temporal superframe segmentation for user videos and producing informative summaries from them using a 0/1- knapsack optimization. In this method,the video is segmented using a "superframe" segmentation, tailored to raw videos. Then, visual interestingness per superframe using a set of low-, mid- and high-level features is estimated. Based on this scoring, an optimal subset of superframes is selected to create an informative and interesting summary. This method introduced a new benchmark dataset, SumMe, it contains 25 raw or minimally edited user videos covering holidays, events and sports. The model has an average performance of 52% relative to the upper bound.

Yale Song et. al. (2015) [14] proposed TVsum(Title-based Video Summarization) an unsupervised video summarization framework that uses the video title to find visually important shots. In this method, title-based image search results are used to generate a summary by selecting shots that are the most relevant to, and representative of, canonical visual concepts shared between the given video and images. The framework consists of four modules: shot segmentation, canonical visual concept learning, shot importance scoring, and summary generation. This method introduced a new benchmark dataset, TVSum50, that contains 50 videos and their shot-level importance scores annotated via crowdsourcing. This approach achieved a mean pairwise F1 score of 0.2655; Interestingness achieved 0.2345, Web Image Prior achieved 0.2403, and LiveLight achieved 0.2438.

Aidean Sharghi et. al. (2016) [15] proposed a method called Sequential and Hierarchical Determinantal Point Process (SH-DPP) for query based Video summarization. The shots are extracted from the video and included in the summary based on the user query and importance in the context of the video, jointly. This was the first work on query-focused video summarization.

Mayu Otani et. al. (2016) [16] proposed to use deep video features that can encode various levels of content semantics. For this, a deep neural network that maps videos as well as descriptions to a common semantic space was designed and jointly trained with associated pairs of videos and descriptions. To generate a video summary, the deep features from each segment of the original video was extracted and a clustering-based summarization technique was applied to them. Their video summaries achieved 58.8% of the average score of manually-created video summaries, while VGG-based got 40.8%.

Ke Zhang et. al. (2016) [17] proposed a novel subset selection technique that leverages supervision in the form of human-created summaries to perform automatic keyframe-based video summarization. In this method, a determinantal point process (DPP) is used for subset selection. Then, frame based visual similarity is measured and idealised summarization kernels are defined. The kernels are transferred from training to testing videos and then the summaries are extracted. This model shows better f score fon\r the Open Video Project (OVP), the YouTube dataset and the Kodak consumer video dataset than the VSUMM model.

Mohaiminul Al Bahian et. al. (2017) [18] proposed a Convolutional Neural Network to extract shots for summarization. The weights and biases of the CNN are trained extensively through off-line processing, so that it can provide the importance of a frame of an unseen video almost instantaneously. This method is better than the commonly referred feature-based methods for estimating the shot importance in terms of mean absolute error, absolute error variance, and relative F-measure.

Kaiyang Zhou et. al. (2018) [19] proposed a Deep summarized network(DSN) to summarize videos. In this method the frames are analysed sequentially and given probability for which predicts if the frame will be selected. This method is trained using end-to-end reinforcement learning based framework,wherea novel reward function is designed that jointly accounts for diversity and representativeness of generated summaries and does not rely on labels or user interactions at all. During training, the reward function analyses how diverse and contextual the summarised videos are, , while DSN strives for earning higher rewards by learning to produce more diverse and more representative summaries. Since labels are not required, this method is fully unsupervised. In unsupervised approaches, It can be seen that DR-DSN outperforms the other unsupervised approaches on both datasets by large margins. On SumMe, DR-DSN is 5.9% better than the state-of-the-art, GANdpp. On TVSum, DR-DSN substantially beats GANdpp by 11.4%. In supervised method, In terms of LSTM-based methods, the DR-DSNsup beats the others, i.e., Bi-LSTM, DPPL STM and GANsup, by 1.0% □ 12.0% on SumMe and 3.2% □ 7.2% on TVSum, respectively.

Jiri Fajtl et. al. (2019) [20] proposed a pure attention, sequence to sequence network VASNet for video keyshots summarization. In this method, there is a soft, self-attention and a two layer, fully connected network for regression of the frame importance score. First layer has a ReLU activation followed by dropout and layer normalization. Second layer has a single hidden unit with sigmoid activation. Kernel Temporal Segmentation (KTS) is used to detect scene change points. This method has an improvement by 0.7% and 1% in the canonical and augmented settings respectively when evaluated using the TVSum dataset and the improvement is by 12% and 11% in the canonical and augmented settings respectively when evaluated using the SumMe dataset.

Lebron Casas et. al. (2019) [21] proposed a deep learning architecture with two LSTMs(vsLSTM and dppLSTM) and an attention mechanism for video summarization. The LSTMs helps in generating a representative summary of an input video by extracting the most relevant segments.The vsLSTM consists of bidirectional chains of LSTM units. The dppLSTM combines the vsLSTM network with a determinantal point process (DPP) to model pairwise repulsiveness among video frames. Furthermore, an attention mechanism was incorporated to learn how the user's interest evolves along the video duration. The proposed model produces a better F score compared to other LSTM models, but has a higher computational cost than the other models.

Bin Zhao et. al. (2017) [22] proposed a hierarchical recurrent neural network(H-RNN) for video summarization, with two layers - where the first layer is a LSTM and the second layer is a bi-directional LSTM. The first layer is used to encode short video subshots cut from the original video, and the final hidden state of each subshot is input to the second layer for calculating its confidence to be a key subshot. It is tested on the datasets - Combined and VTW. The proposed H-RNN model performs better than other RNN models like vsLSTM and dppLSTM.

Pinelopi Papalampidi et. al. (2020) [23] proposed a model that identifies the key events in a movie that describe its storyline(turning points) by building a sparse movie graph that represents relations between scenes and is constructed using multimodal information. In this method initially a dense complete graph with edge weights representing the probability of scenes being adjacent to each other. Then, similarity between two scenes are computed and then a sparse graph is obtained by constructing a k-NN graph. Graph convolutions are performed on top of LSTM states This model produces comparatively lengthy summaries for movies(over 30 mins) compared to other models.

George Pantazis et. al. (2020) [24] proposed a method based on a Generative Adversarial Network (GAN) model pre-trained with human eye fixation saliency information. In this method, saliency maps are generated for each frame using a GAN model and the hue histograms of video frames are calculated to estimate a static score. Then, a spatiotemporal score is computed by calculating the optical flow between saliency maps. A final score is computed by fusing the two scores. The keyframes are selected by considering the local minima of the signal formed by this score over time. This method achieved an average f-measure score of 0.835 where the second best performing method achieved a score of 0.788 when evaluated using the VSUMM dataset.

Table 1 shows the algorithms that were used and the results that were obtained by various authors.

TABLE 1: Summary of Various Video Summarization techniques and methods

Author's Name	**Algorithm/Techniques Used**	**Results**
Daniel DeMenthon, Vikrant Koblaet. al. 1998	Videos represented as curves, Binary curve splitting algorithm	-----
Anastasios D. Doulamis, Nikolaos D. Doulamis et. al. 2000	Colour/motion segmentation algorithm - M-RSST(Multiresolution recursive shortest spanning tree)	Performed better than the logarithmic method by providing a much better description of the visual content compared to it.
Yoshimasa Takahashi, Naoko Nitta et. al. 2004	Temporal compression of videos and Spatial representation of keyframes	------
Jian-quan OUYANG, LI Jin-tao et. al. 2005	Visual Saliency - Summarization based on replays and liveshots	The accuracy and precise of replay scene detection is fairly good, and the recall of identical events detection is also 100%.
G. Evangelopoulos, K. Rapantzikos et. al. 2008	Audio Saliency - captured by signal modulations Visual saliency measured using spatio temporal attention model	The videos received 79.4%, 66.7%, 54.4 % rating for enjoyability and 85.5%, 77.7%, 65.9% rating for informativeness when tested on three videos of the MUSCLE movie database.
Robert Laganière, Raphael Bacco et. al. 2008	Spatio Temporal	The summaries were judged to have a very pleasant rhythm in the framework of TRECVID 2008.
Jitao Sang, Changsheng Xu 2010	Substory discovery, Content attention analysis	This method achieved better enjoyability scores than the method proposed by[25].
Sandra Eliza Fontes de Avila, Ana Paula Brandão Lopes et. al. 2011	Visual Saliency, k-means clustering	Proposed the benchmark VSUMM Dataset and the model produced a mean accuracy rating of 0.85 against the custom user selected summaries.
Irfan Mehmood, Muhammad Sajjad et. al. 2015	Deep semantic feature extraction	Comparison with Non-Visual Attention Based Video Summarization Techniques gives 92.12 and 89.8 based on informativeness and enjoyability.
Karim M. Mahmoud, Mohamed A. Ismail et. al. 2013	Color feature extraction using color histogram, Texture feature extraction 2D Haar wavelet transform, DBSCAN clustering	F-measure score of 0.77 for the Open Video Project Dataset.
Bin Zhao, Eric P. Xing 2014	Dictionary generation using group sparse coding	LiveLight outperforms the state-of-the-art summarization method by 8%. Achieve 40 times higher compression without losing semantic meaning.
A. Raventós, R. Quijada et. al. 2014	Audio Visual Descriptors	The proposed system was able to satisfactorily detect over 70% of the total amount of goals of the five soccer matches analyzed.
Michael Gygli, Helmut Grabner et. al. 2014	0/1 - Knapsack optimization	The method has an average performance of 52% which outperformed all the baselines at that time.
Yale Song, Jordi Vallmitjana, Amanda Stent et. al. 2015	Title based Keyframe Extraction	Proposed the benchmark TVSum Dataset and the model produced a mean F1 Score of 0.50 against it.
Aidean Sharghi, Boqing Gong et. al. 2016	Sequential and Hierarchical Determinantal Point Process (SH-DPP)	SH-DPP has a better f score than seqDPP and DPP.
Mayu Otani, Yuta Nakashima et. al. 2016	Video-Description mapping using Deep Neural Network	The video summaries achieved 58.8% of the average score of manually-created video summaries.
Ke Zhang , Wei-Lun Chao et. al. 2016	Subset selection using Determinantal Point Process (DPP)	The method produces a f-score of 82.3, 76.5, 61.8, 30.7, 40.9 for the Kodak,

		Open Video Project, Youtube, MED, SumMe benchmark datasets respectively.
Mohaiminul Al Nahian, A. S. M. Iftekhar et. al. 2017	Convolutional Neural Network	The method produces a relative F-measure score of 0.722, mean absolute error of 0.3212, absolute error variance of 0.0572 for the TVSum dataset.
Kaiyang Zhou, Yu Qiao et. al. 2018	Deep Summarized Network (DSN) trained using Reinforcement learning	The Area-Under-the-Curve (AUC) of average F1-scores were 15.87 for Π temporal filter and 15.43 for Gaussian temporal filter for a consumer grade egocentric videos dataset.
Jiri Fajtl, Hajar Sadeghi Sokeh et. al. 2019	Attention	The method produces a pairwise F-score of 53.8 for the TvSum dataset and 31.8 for the SumMe dataset.
L. Lebron Casas, E. Koblents 2019	Long Short Term Memory (LSTM), Determinantal Point Process	The method produces a F1 score of 43.2 and 63.1 for the SumMe and TVSum Dataset respectively.
Bin Zhao, Xuelong Li et. al. 2017	Hierarchical Recurrent Neural Network(H-RNN)	H-RNN has an f-score of 0.451 better than VSUMM, livelight and other popular methods on a combined dataset. H-RNN has 0.487 f-score better than the LSTM methods on VTW dataset.
Pinelopi Papalampidi, Frank Keller et. al. 2020	Sparse Graph Construction	The Summaries received a rating of 3.02 from the users on a scale of 1 to 5, with 5 being the highest
George Pantazis, George Dimas, Dimitris K et. al. 2020	Generative Adversarial Network (GAN)	The method produces a F-measure score of 0.835 for the VSUMM Dataset

REFERENCES

1. Daniel DeMenthon, Vikrant Kobla, David Doermann. **Video Summarisation by curve Simplification**, *MM98: The Sixth ACM International Multimedia Conference*, pp. 211-118,September 1998.
2. Anastasios D. Doulamis, Nikolaos D. Doulamis, Stefanos D. Kollias. **A fuzzy video content representation for video summarization and content-based retrieval**, *Signal Processing, Volume 80, Issue 6*, pp.1049-1067, June 2000.
3. Yoshimasa Takahashi, Naoko Nitta, and Noboru Babaguchi. **Automatic Video Summarization of Sports Videos Using Metadata**, *PCM '04, November 2004*, Volume.Part.II, pp. 272–280.
4. Jian-quan OUYANG, LI Jin-tao, ZHANG Yong-dong. **Replay Scene Based Sports Video Abstraction**,*Second International Conference, FSKD 2005, Changsha, China*, pp. 45-47, August 2005.
5. G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, Y. Avrithis. **Movie summarization based on audiovisual saliency detection**, *15th IEEE International Conference on Image Processing (ICIP-2008)*, pp.2528-2531, October 2008.
6. Robert Laganière, Raphael Bacco, Arnaud Hocevar, Patrick Lambert, Grégory Païs, Bogdan E. Ionescu. **Video summarization from spatio-temporal features**, *Proceedings of the 2nd ACM TRECVid Video Summarization WorkshopTVS '08*, October.2008, pp.144–148.
7. Jitao Sang, Changsheng Xu. **Character-Based Movie Summarization**, *MM '10: Proceedings of the 18th ACM international conference on Multimedia*, October.2010, pp.855–858.
8. Sandra Eliza Fontes de Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr., Arnaldo de Albuquerque Araújo. **VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method**, *Pattern Recognition Letters*, Volume 32, Issue 1, January 2011, pp.56–68.
9. Irfan Mehmood, Muhammad Sajjad, Sung Wook Baik. ***Visual attention based extraction of semantic keyframes***, Advances in Information Science and Applications - Volume 1, 2015.
10. Karim M. Mahmoud, Mohamed A. Ismail, and Nagia M. Ghanem. **VSCAN: An Enhanced Video Summarization Using Density-Based Spatial Clustering**. *Image Analysis and Processing – ICIAP* 2013, pp 733-742
11. Bin Zhao, Eric P. Xing. **Quasi Real-Time Summarization for Consumer Videos**, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2513-2520, September 2014.
12. A. Raventós, R. Quijada, Luis Torres, Francesc Tarrés. **Automatic Summarization of Soccer Highlights Using Audio-visa,** *SpringerPlus, Vol 4, November 2014.*
13. Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Luc Van Gool. **Creating Summaries from User Videos**, *ECCV 2014: Computer Vision*, 2014, pp.505-520.
14. Yale Song, Jordi Vallmitjana, Amanda Stent, Alejandro Jaimes. **TVSum: Summarizing Web Videos Using Titles**, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5179-5187, June 2015.
15. Aidean Sharghi, Boqing Gong, Mubarak Shah. **Query-Focused Extractive Video Summarization,** *European Conference on Computer Vision*, 2016, pp.3-19.
16. Mayu Otani, Yuta Nakashima , Esa Rahtu, Janne Heikkil¨a , Naokazu Yokoya**. Video Summarization using Deep Semantic Features**, *ACCV 2016: Computer Vision*, 2016, pp.361-377.
17. Ke Zhang , Wei-Lun Chao, Fei Sha, Kristen Grauman. **Summary Transfer: Exemplar-based Subset Selection for Video Summarization**, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1059-1067, June 2016.
18. Mohaiminul Al Nahian, A. S. M. Iftekhar, Mohammad Tariqul Islam, S. M. Mahbubur Rahman, Dimitrios Hatzinakos. **CNN-Based Prediction of Frame-Level Shot Importance for Video Summarization**, *International Conference on New Trends in Computing Sciences (ICTCS)*, 2017, pp.24-29.
19. Kaiyang Zhou, Yu Qiao, Tao Xiang. **Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward**, *arXiv:1801.00054*, 2018.
20. Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. **Summarizing Videos with Attention,** *ACCV 2018: Computer Vision*, June 2019, pp.39-54.
21. L. Lebron Casas, E. Koblents. **Video Summarization with LSTM and Deep Attention Models**, *Proc. International Conference on MultiMedia Modeling. MMM*, Greece, 2019, pp.67-79.
22. Bin Zhao, Xuelong Li, Xiaoqiang Lu. **Hierarchical Recurrent Neural Network for Video Summarization**, *In Proceedings of the 25th ACM international conference on Multimedia (MM '17). Association for Computing Machinery*, New York, 2017, pp.863–871.
23. Pinelopi Papalampidi, Frank Keller, Mirella Lapata. **Movie Summarization via Sparse Graph Construction**, *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, February 2020.
24. George Pantazis, George Dimas, Dimitris K. Iakovidis. **SalSum: Saliency-based Video Summarization using Generative Adversarial Networks**, *arXiv:2011.10432*, 2020.
25. Ma, Yu-Fei & Hua, Xian-Sheng & Lu, Lie., **A generic framework of user attention model and its application in video summarization**, *Multimedia, IEEE Transactions*, pp.907-919, November, 2005.