2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP) | 978-1-6654-3277-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCSP52374.2021.9465347

Survey and Comparison of Video Summarization Techniques

Anil Singh Parihar, Ritvik Mittal, Prashuk Jain and Himanshu
Dept. of Computer Science
Delhi Technological University
New Delhi, India
{parihar.anil, mittalritvik.rm, himanshubhushan775, prashuk156}@gmail.com

***Abstract*—Video summarization is a keenly intellective video compression technique to select a subset of keyframes or keyshot which are combined to represent shorter and compendious summary of the original input video without losing the contextual semantics of the same. In context of summary generated, it can be divided into static (static storyboard) and dynamic (video skimming) video summarization. Supervised ,unsupervised and reinforcement learning techniques have been introduced in the literature of video summarization. Earlier development in this field marked the introduction of various unsupervised techniques that used handcrafted heuristics to select mutually independent keyframes. In recent years, supervised techniques and deep reinforcement learning techniques have been introduced that models the structural semantics of the original video to generate summaries using frame - level ground truth generated by humans, so that we get the result as close as possible to the human understanding of the video. This paper aims to introduce different architectures proposed for video summarization and provides a qualitative and quantitative comparison of these methods.**

***Index Terms*—video, summarization, survey, quantitative, qualitative**

I. Introduction

With the advent of high speed Internet and low storage cost, the amplitude of data has incremented dramatically and the major part of it is represented in the form of visual data or videos. With such humongous magnitude of data, we require efficacious techniques and implements to take these videos and represent them in a more compact and concise way so that these can be utilized for sundry applications like video browsing, video surveillance analysis, egocentric video analysis to study human comportment, medical video analysis like endoscopy analysis, etc. These requisites magnetized researchers and technologists and paved the way for incipient research areas like video summarization.

Video summarization is usually considered to be a selection optimization problem and models are built to select such frames or shots which store the highest representative value amongst the frame. There are various representations that have been considered as an output to video summarization tasks like video synopsis [9], montages [1, 10, 11] and storyboards [12, 13]. This computer vision task of video summarization can be divided into precisely two subgroups - static video summarization and dynamic video summarization. Static video summarization works at the level of frames and considers it as a keyframe selection problem. In dynamic video summarization, keyshots are selected taking into account the domain representation of the frames. The major advantage of dynamic video summarization is that it introduces motion as well as audio element of the video into the summary. Audio and motion help to improve the expressiveness of the summary and keep the viewer more engaged and intrigued. But dynamic video summarization also needs synchronization of the keyshots and hence engenders bottlenecks on the flexibility of the output summary.

Initially research work was majorly carried out in static video summarization but now with the amplitude of resources available, researchers are moving towards producing models that work on dynamic video summarization. In recent years, deep learning has been extensively used for the task of video summarization. In this context, video summarization can be subdivided into supervised and unsupervised tasks. Earlier, usually unsupervised models were introduced for the task that leveraged low level feature extraction like color distribution, pixel values, etc and clustering techniques to select keyframes. But now supervised models are being designed for the task which use human annotated video dataset to model architecture that could generate summary as close as possible to human interpretation of the video. Due to the subjective nature of annotations, it is very difficult to design an established evaluation metric. Nevertheless, F-score is a generally accepted evaluation metric used for quantitative analysis of various models.

In this paper a detailed survey is provided for various models that have been introduced for video summarization and a comparative analysis is generated including both quantitative and qualitative aspects.

978-1-6654-3277-1/21/$31.00 ©2021 IEEE

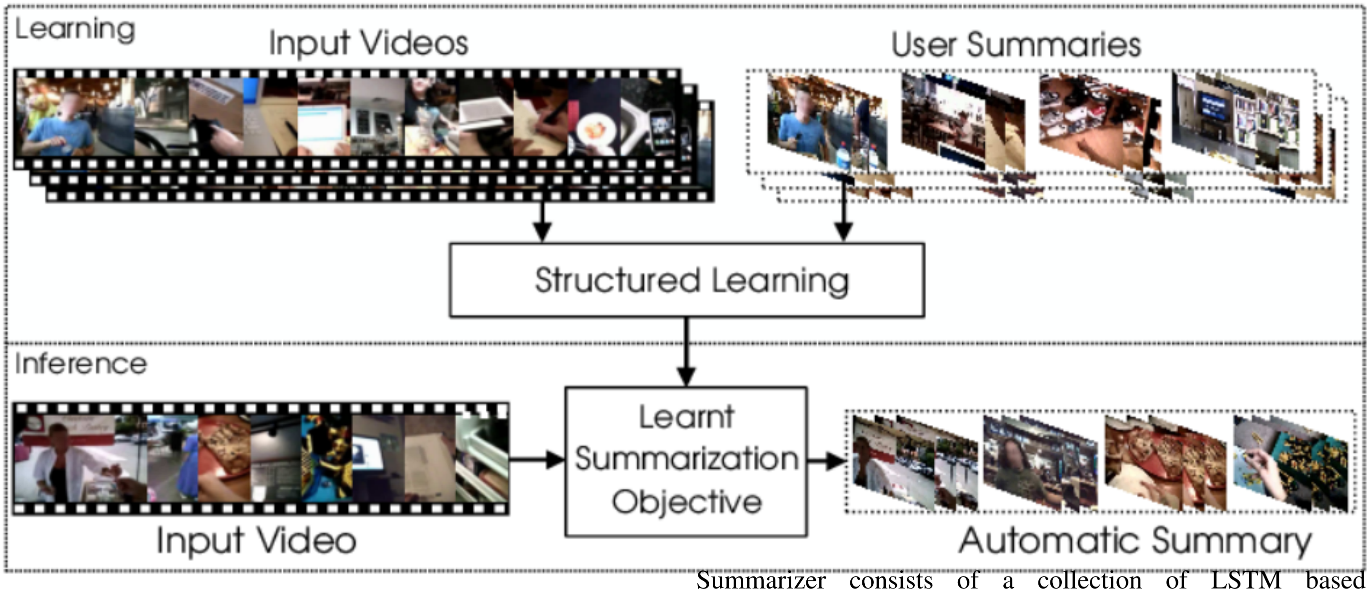

Fig. 1. Standard operation of Video Summarization Models [18]

II. Related Work

Sandra Eliza Fontes de Avila et al. [1] proposed an efficient unsupervised video summarization approach known as VSUMM. It extracts features from some sample video frames in the HSV color space and then applies K-means clustering on the extracted features to group them into different clusters. Then a representative frame is chosen for each cluster and the summary is generated. VSUMM also introduces a new evaluation technique known as CUS (Comparison of User Summaries). In CUS, users create a summary for the input video and that summary is compared with the summary created with the model. A frame in the user summary is considered similar to that created by the model if the Manhattan distance between them is less than some predefined value delta and consequently the accuracy and the error is calculated.

Ke Zhang et al. [2] leveraged the adeptness of LSTMs in extracting the structural and contextual knowledge from a sequence towards video summarization. A BiLSTM network is proposed to model the structure of the video considering both the directions. The network is trained in a supervised setting against the user generated summaries to distillate the semantics of how a user perceives summarizing a video. One other major contribution was standardization of techniques to deal with unavailability of user annotated video summaries.

Behrooz Mahasseni et al. [3] introduced a unique adversarial learning based model for unsupervised video summarization. The architecture focuses on learning a summarizer network to generate video summaries which are evaluated against the original video using a discriminator network. Summarizer consists of a collection of LSTM based network modules for frame selection, encoding and decoding. Discriminator is modeled as a LSTM network which takes as input original and generated summaries and tries to differentiate them. The summarizer is trained to maximally confuse the discriminator thus minimising the difference between original and generated summary. LSTMs are used as they can model long range semantic dependencies between the temporally ordered input frames.

Mrigank Rochan et al. [4] casts video summarization as a binary sequence labeling problem and establishes a relationship with a far more researched topic of semantic segmentation. [4] thus leverages the use of fully convolutional networks [9] for video summarization. The network architecture is in the form of an encoder decoder network. Stack of convolutions and pooling layers aggregate the contextual information between the frames and form the encoder. The reduced temporal dimensions are then recovered via upsampling of the encoder outputs. The whole architecture is then realised as an end-to-end convolutional network. Main intuition is repeated convolutions can effectively capture range and structural dependencies much like a LSTM but are easier to train.

Kaiyang Zhou et al. [5] proposed the first reinforcement learning based unsupervised video summarization technique. In this technique the reward function calculates how diverse and representative the selected frames are for the input video. The diversity function was proposed in such a way that two distant similar frames are considered different as they can be important for the storyline construction. Diversity value is calculated by taking the pairwise dissimilarity among selected video frames. No labelling of frames is required for this approach however the model is extended by considering log probability of selecting the annotated keyframes when labels are present. The paper proposed a DSN (Deep Summarization Network) which predicts a probability for each frame which represents the probability that the frame is selected. DSN is an encoder decoder network where encoder is a CNN network which extracts features from the input video and the decoder is a Bi-LSTM network that produces the probabilities for frame selection. Pre-trained GoogleNet is used as the CNN encoder network and only decoder part is trained.

Zhong Ji et al. [6] proposed an attention based encoder decoder network where the encoder is a Bi-LSTM network and the decoder is an attention based LSTM network. The proposed model is named AVS (Attentive encoder decoder networks for video summarization). The input to AVS is the video frames and the output of the AVS is the frame scores. A key shot selection model is then applied on the frame level importance scores to generate a shot level video summary. The proposed model improves upon previous Bi-LSTM encoder decoder approaches by allowing the context vector, that flows from encoder to decoder, to be of variable size. In previous Bi-LSTM encoder decoder networks, no matter what kind of output shots/frames were to be predicted, all the shots/frames in the input video series had the same importance. Those methods risked missing much of the temporal context underlying the video because of this indiscriminate averaging of all the frames. The proposed model gives different importance to different frames using the attention based decoder network thus trying to generate the summary like a human does.

Jiri Fajtl et al. [7] proposed a supervised learning approach with a Soft Attention based Simple Neural Network as it hypothesizes that video summarization is rather a subjective task. The paper proposed a new soft attention based simple network for sequence to sequence transformation and demonstrates that the proposed network is better and is less complex than the already used Bi-LSTM based encoder decoder network with soft attention. The proposed network is a combination of a soft attention network and a regressor network. In some Bi-LSTM based encoder decoder networks for video summarization, the state that is passed from encoder to decoder network is of fixed length, so if the input to the network is a sequence of longer length then the state passed from encoder to decoder will have more information loss. The proposed network doesn't suffer from such losses.

Jungin Park et al. [8] emphasizes on modeling the semantic relationships between the frames with respect to the story of the whole video. It drifts from other methods in casting the problem as graph modeling problem in which individual frames represent nodes while edges represent the relationship between these nodes. An initial graph constructed using frame level features and represented by an adjacency matrix is iteratively refined using Graph Convolution outputs as adapted feature representations for each step. This architecture has an effect that the neighbourhood of each node is updated depending on the effective context at each step. The final frame features and relations are fed to a summary GCN to classify whether it is part of the generated summary. Network can be trained in supervised as well as in unsupervised setting.

III. Datasets

Supervised training requires user generated summaries of training videos. TVSum [16] and SumMe [15] are two annotated datasets which contain videos summarised by multiple users. These datasets are the benchmark for testing video summarization models. TVSum has 50 videos with categories ranging across news, documentaries, etc. and video length 1-5 minutes. It provides user generated summaries as frame level importance values. SumMe dataset has 25 videos of various events across holidays, sports etc. ranging from 1.5 to 6.5 minutes in length. User generated summaries are key shot based (set of frame intervals). OVP(open video project) [1], [17] dataset has 50 videos covering genres such as historical, educational, etc. These are 1-4 minutes in length.The YouTube [1] dataset contains 39 videos distributed among news, sports,commercials, etc. These are 1-10 minutes in length. Both OVP and YouTube dataset provide multiple user summaries as key-frames. Since different datasets provide user annotated summaries in different formats, we follow [2] for pre-processing data for training of supervised models and evaluation.

IV. Comparative Analysis

We follow the commonly used evaluation metric [2] and use F-Scores for quantitative comparison. It is most commonly used metric across the work we have considered and hence provide for fair analysis. F score is based on precision(P) and recall(R). Given G as ground truth summary and M as machine generated summary:

$$P = G \cap M / |M| \qquad R = G \cap M / |G| \tag{1}$$

where $|.|$ indicates summary length. F-Score is computed as:

$$F = (2P * R)/(P + R) * 100\% \tag{2}$$

F scores on augmented [2] setting of the datasets where training is performed on augmentation of OVP, YouTube, X1, 80% of X2 and testing on 20% of X2 are given. X1 and X2 being TVSum and SumMe interchangeably. Table I shows the quantitative analysis based on F-score of the various research models we have studied while Table II presents a brief qualitative analysis on these techniques.

TABLE I
F scores of Analysed models

Paper	SumMe	TvSum
[1]	33.5	48.6
[2]	42.9	59.6
[3]	43.6	61.2
[4]	51.1	59.2
[5]	43.9	59.8
[6]	46.1	61.8
[7]	51.1	62.4
[8]	52.9	65.8

TABLE II
QUALITATIVE ANALYSIS OF MODELS

Author	Approach	Advantages	Drawbacks
Sandra Eliza Fontes de Avila et al. [1]	In this approach, Hue color histograms as feature desciptors and K Mean clustering selects clusters which are visually diverse	1. Colour histograms are used as low level feature descriptors which are robust to small changes in camera position 2. This approach produces good results using very less resources as the time and computer resources required to form clusters is very less compared to neural net architectures.	1. Clustering Algorithm does not take into account the temporal order 2. Colour histograms do not take into account the orientation or the colour spread in the frame. Thus an image with the same colours distributed very differently gets considered similar to an image with same colours distributed differently
Ke Zhang et al. [2]	The model uses BiLSTM architecture. Output from these LSTM layers is coupled with visual frame feature into multi-layer perceptron to yield either binary frame label or frame level importance score at each temporal step.	LSTMs model the variable temporal dependencies effectively because of persistent past knowledge.	1. The computations of recurrent models cannot be parallelised hence leading to a waste of computing resources. 2. Poor performance on modelling dynamically changing scenarios as structural order is not maintained.
Behrooz Mahasseni et al. [3]	The architecture learns a keyframe selector LSTM by using variational autoencoder to generate corresponding summary which is then fed to confuse the discriminator in an adversarial setting.	1. Unsupervised Approach thus no need for annotated user data which is hard to obtain. 2. Learns Intermediate representations maybe useful for other applications. 3. Adversarial feedback maybe more useful than user ground truths in some cases. 4. Regularization can be introduced at various instances thus focusing on varying semantics.	1. GANs are notorious to train both in time and memory. 2. Subtle changes which may be important to users are hard to capture as focus is on achieving global representation. 3.User semantics in generating a video summary are not captured.
Mrigank Rochan et al. [4]	A fully convolutional model is adapted for video summarization.	1. Convolution models support parallel computation as they are not dependent on the previous results. This provides for efficient training compared to LSTM approaches. 2. Compared to LSTMs, CNNs can model whole range at much smaller depth allowing for high level context aggregation earlier in the network. In LSTMs, the last node is only considered once.	1. Repeated down-sampling of the inputs by the stack of convolution layers results in loss of resolution and low level semantics thus biasing the output towards contextual knowledge only disregarding local knowledge. 2. Upsampling by a large factor results in spreading of a small number of values to a large space thus affecting uniqueness and rendering many nodes similar.
Kaiyang Zhou et al. [5]	Reinforcement learning framework for training of Encoder Decoder Architecture. Rewards are based on diversity and representativeness of generated summary.	1. No requirement of annotated data for training. 2. Directly modelling diversity and representativenss provides better results than reconstruction loss [3].	1. Focus on diversity introduces problems in case of subtle or slow changes. 2. LSTMs cause inefficient training. 3. Does not include any reward towards long temporal dependencies.
Zhong Ji et al. [6]	Attention based encoder decoder network where the encoder is a Bi-LSTM network and the decoder is an attention based LSTM network.	Attention mechanism provides attention weights associated with individual encoder hidden states. Decoder output at a single instance can learn to combine different hidden units through these weights resulting in rich structural knowledge. This is particularly beneficial against encoded vector representation of inputs which provide fixed knowledge. The latter technique is common in other LSTM encoder decoder techniques [2], [3] and results in loss of temporal structural especially in case of large inputs.	The model is computationally expensive to train as both the encoder and decoder are LSTM based networks. More memory and time requirements to compute attentions at each time step based on all encoder outputs.
Jiri Fajtl et al. [7]	The model is based on evaluating a soft, global attention layer to evaluate a context vector using attention weights. Context vector along with residual sum is then used for frame level score regression.	1. Provides the advantage of attention mechanism as in [6] while being computationally efficient and easier in implementation. Does not involve any sequential processing and just 2 fully connected layers. 2. Proposed model prevents compression losses as in encoder decoder architectures by accessing the input features directly. 3. Impressive results without taking temporal structure into account.	1. No consideration of temporal semantics. 2. Slight increase in accuracy on the TvSum dataset as compared to the SumMe indicates global attention is not desirable for long videos. Hence local attention is required.
Jungin Park et al. [8]	Initialised graph from frame features affinity is refined recursively using Graph Convolutional Networks which follow aggregating features from previous steps and updation of adjacency matrix.	Frames are related to each other following exploration of similarities among themselves and also through global context as the graph is refined. The limited range of convolutional and recurrent operations cannot model these dependencies to this extent. Hence this model effectively captures frame level relationships.	1. Distance between frames is not accounted and temporal structure is not modeled. 2. Iterative refinement by computing pairwise scores between frames may lead to performance bottlenecks in case of long videos.

V. Conclusion

We have presented different architectures published for video summarization. For each approach, we first introduce the theory and then analyse the architecture and results to furnish a lucid comparison. Through this work, we provide a base for exploring diverse video summarization techniques. We hope this study facilitates current apprehension and future advancement in video summarization.

Our major aim has been to exhibit diverse methodologies for video summarization. To this end, we could not consider some of the other excellent work that has been published for the scope of this study.

References

[1] S. E. F. De Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araujo. Vsumm: A mechanism designed to produce ´ static video summaries and a novel evaluation method. Pattern Recognition Letters, 32(1):56–68, 2011.

[2] Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: ECCV (2016)

[3] Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: CVPR (2017)

[4] Rochan, M., Ye, L., Wang, Y.: Video summarization using fully convolutional sequence networks. In: ECCV (2018)

[5] Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: AAAI (2018)

[6] Ji, Z., Xiong, K., Pang, Y., Li, X.: Video summarization with attention-based encoder-decoder networks. arXiv preprint arXiv:1708.09545 (2017)

[7] Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D., Remagnino, P.: Summarizing videos with attention. In: ACCVW (2018)

[8] Park, Jungin et al. "SumGraph: Video Summarization via Recursive Graph Modeling." Lecture Notes in Computer Science (2020): 647–663. Crossref. Web.

[9] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)

[10] Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1971–1984, Nov 2008.

[11] H. wen Kang, Y. Matsushita, X. Tang, and X. quan Chen. Space-time video montage. In CVPR, pages 1331–1338, 2006.

[12] M. Sun, A. Farhadi, B. Taskar, and S. Seitz. Salient Mon- tages from Unconstrained Videos, pages 472–488. Springer International Publishing, Cham, 2014.

[13] D. B. Goldman, B. Curless, S. M. Seitz, and D. Salesin. Schematic storyboarding for video visualization and editing. ACM Transactions on Graphics (Proc. SIGGRAPH), 25(3), July 2006.

[14] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In CVPR, pages 1346–1353, 2012.

[15] Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: ECCV. (2014)

[16] Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: CVPR. (2015)

[17] : Open video project. http://www.open-video.org/

[18] Gygli, Michael et al. "Video summarization by learning submodular mixtures of objectives." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 3090-3098.