VIDEO SUMMARIZATION USING DEEP NEURAL NETWORKS: A SURVEY

Presented by: Somya Mishra

MS SE, Data Science, SJSU

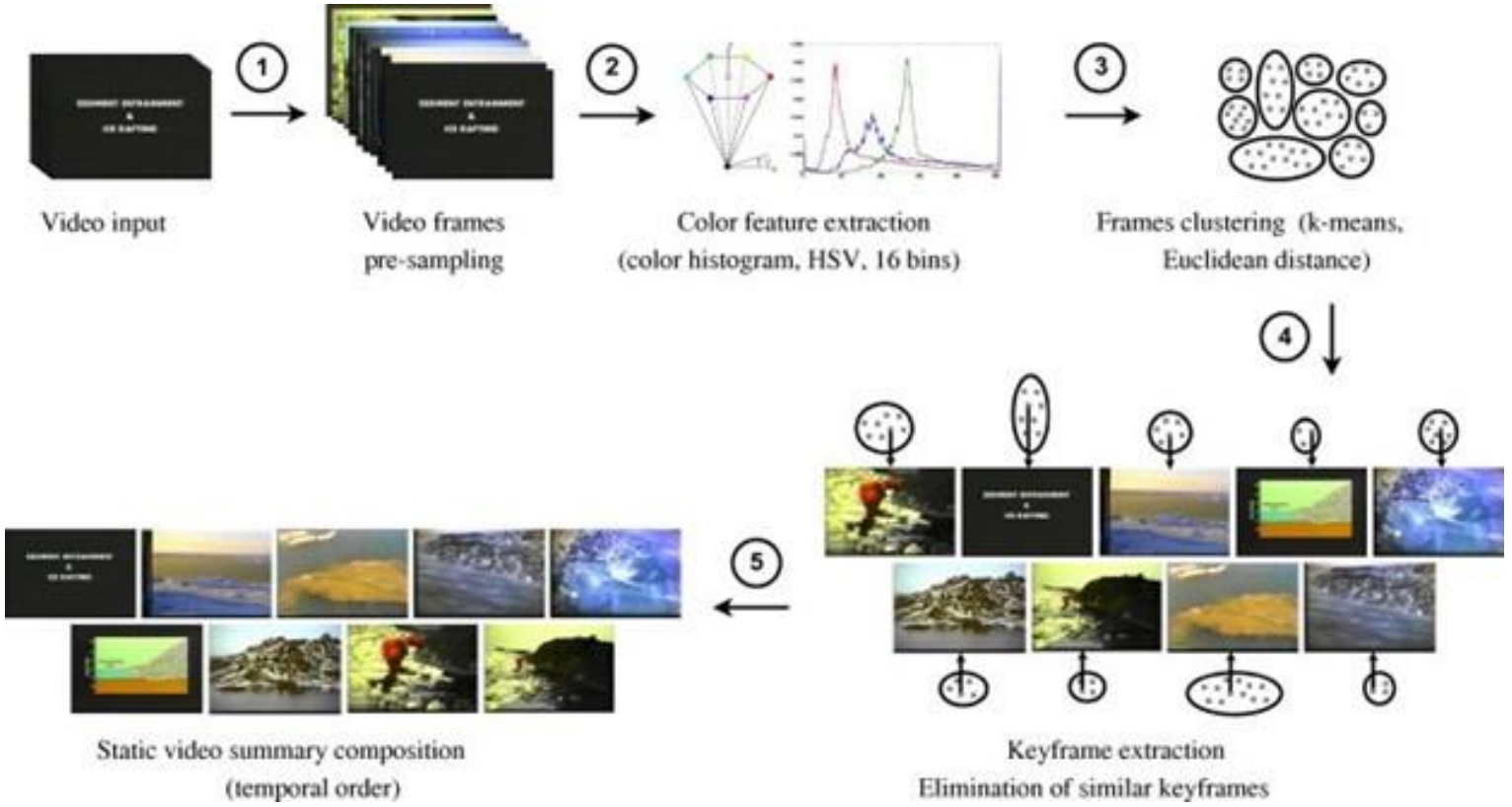
Based on *Video Summarization Using Deep Neural Networks: A Survey*, *published in January 2021*

Authors: *Evlampios Apostolidis | Eleni Adamantidou | Alexandros I. Metsai | Vasileios Mezaris | Ioannis Patras*

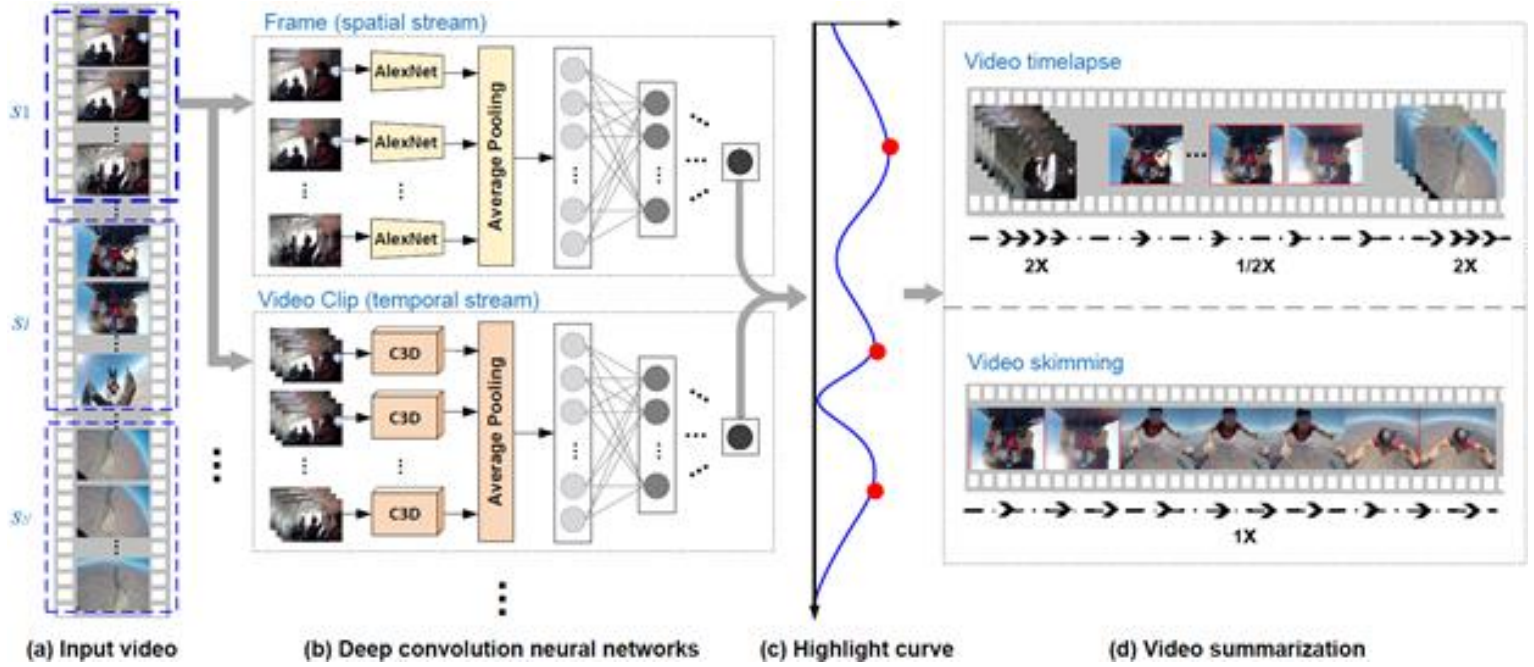
WHAT IS VIDEO SUMMARIZATION?

- Process of shortening videos by selecting frames capturing most informative parts
- Generating a concise and complete synopsis
- Synopsis typically produced in two forms:
  a) Static video summarization: Set of video key-frames (a.k.a video storyboard)
  b) Dynamic video summarization: Set of video key-fragments (a.k.a video skim)

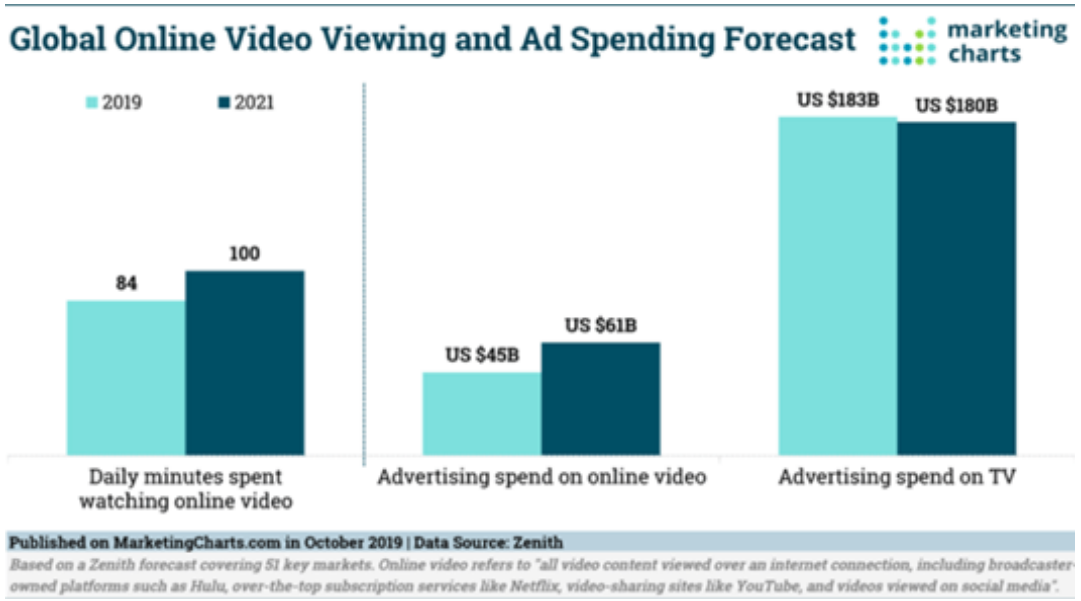
PROCESS OF STATIC VIDEO SUMMARY COMPOSITION



PROCESS OF DYNAMIC VIDEO SUMMARY COMPOSITION



- Constantly growing online video content!
- Rapid advancements in AI & Neural Networks-based learning algorithms
- Over a billion hours of videos watched on YouTube each day!
- Over 400 hours of video content is uploaded every single minute!
- Effective content browsing, enhancing viewing experience
- Increase of consumers' engagement and content consumption



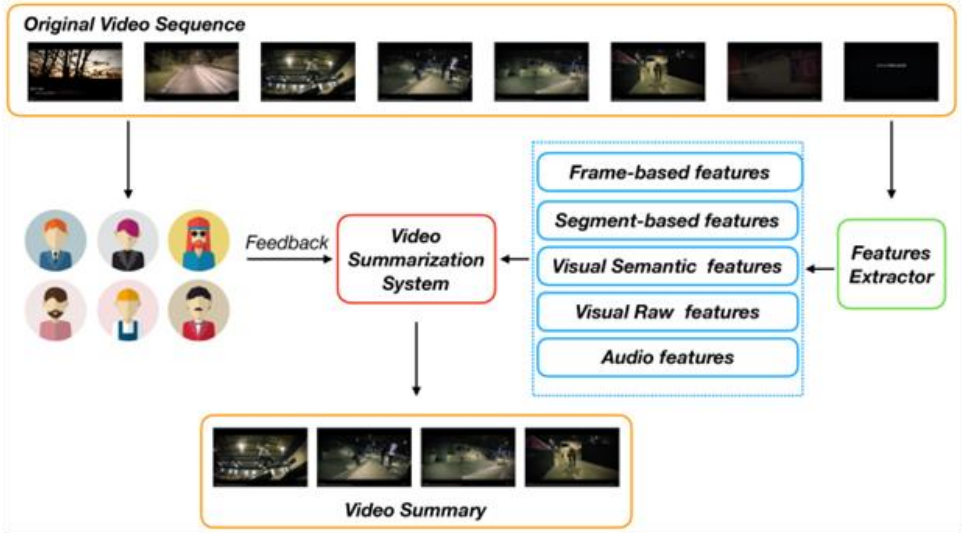
WHY DO WE NEED VIDEO SUMMARIZATION?

Represent visual content of a video in terms of feature vectors

Extract these deep features vectors for all frames or for a subset of them selected via frame sampling

Pass these extracted features through a Deep Summarizer network

Obtain a set of selected video key-frames or a set of selected video key-fragments to form a short video skim as an output.



HOW DOES IT WORK?

BASED ON UTILIZED TYPE OF DATA

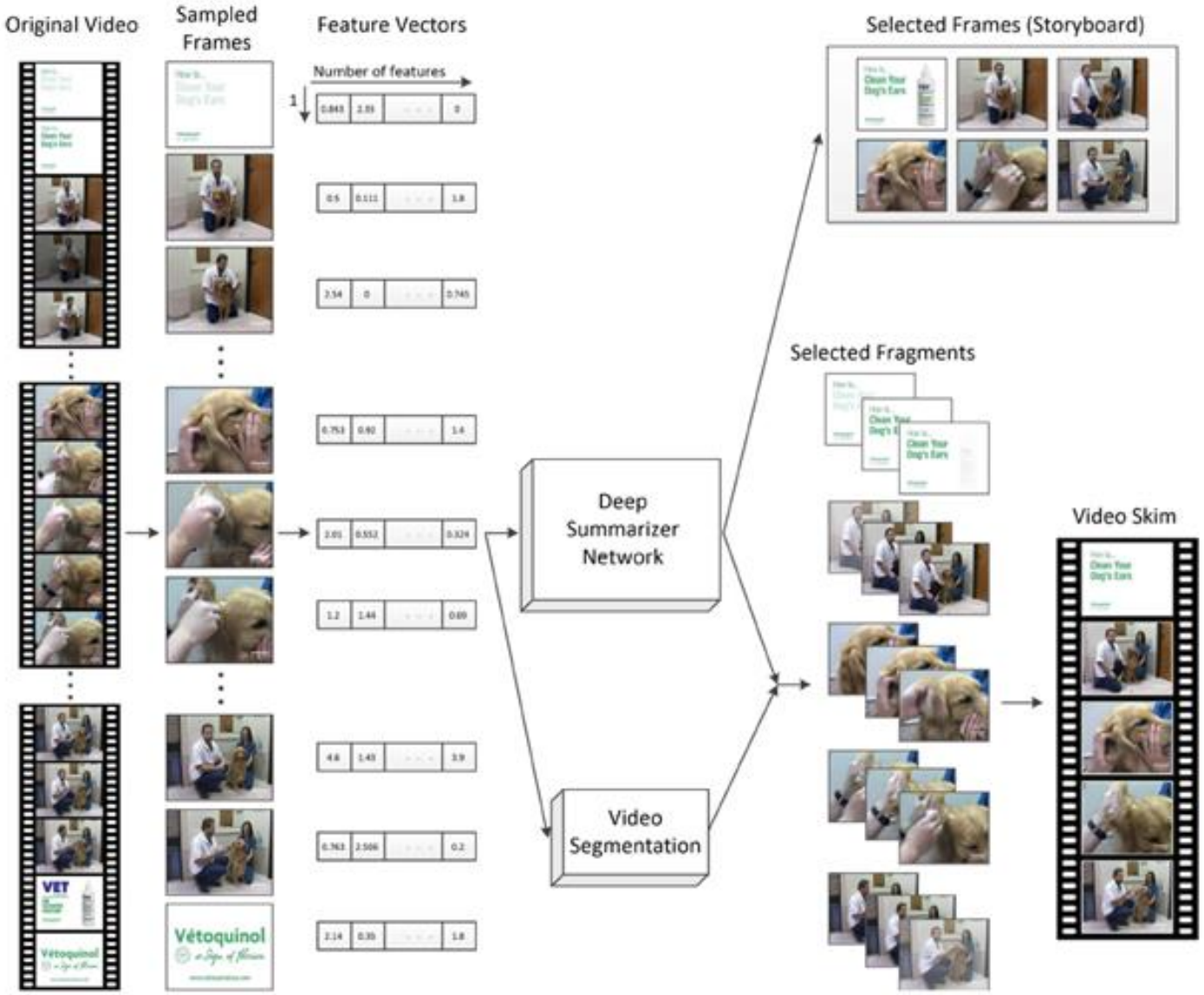
- **Unimodal approaches** where feature extraction is based on visual modality of videos and summarization is learnt in a (weakly-)supervised or unsupervised manner
- **Multimodal approaches** where summarization is learnt in a supervised manner using available textual metadata

BASED ON TRAINING STRATEGY

- **Supervised approaches** dependent on datasets with manually human labeled ground-truth annotations
- **Unsupervised approaches** based on significant collection of original videos for training
- **Weakly-supervised approaches** less-expensive and easy weak labels

TYPES OF DEEP-LEARNING-BASED VIDEO SUMMARIZATION APPROACHES

HIGH-LEVEL REPRESENTATION OF TYPICAL DEEP-LEARNING BASED VIDEO SUMMARIZATION PIPELINE



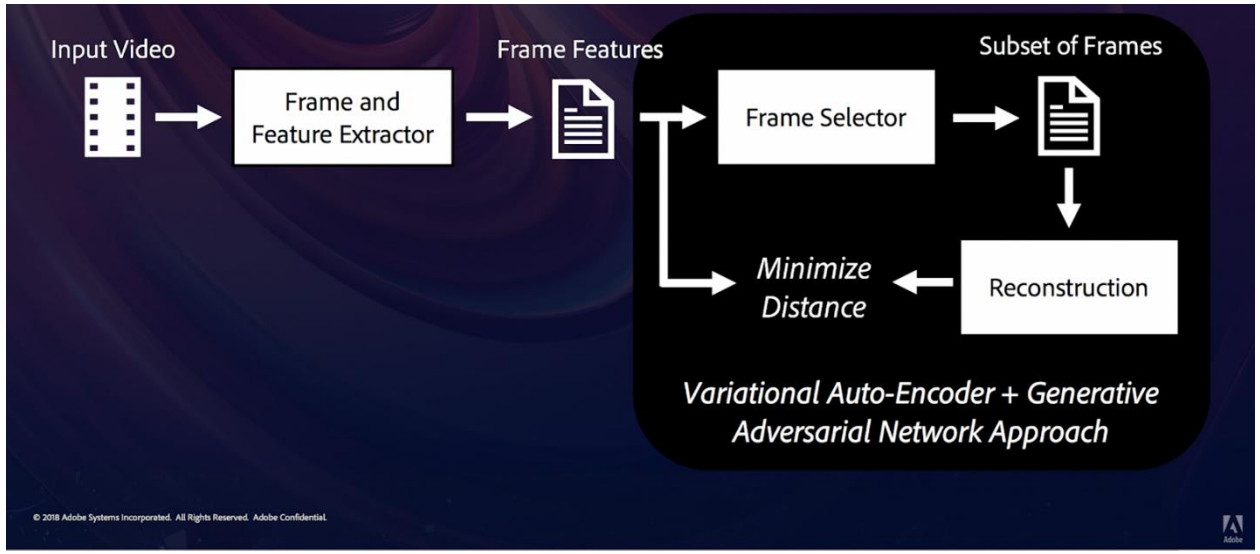
Supervised Video Summarization approach entails:

1. Learn frame importance by modeling temporal dependency among frames
2. Learn frame importance by modeling spatiotemporal structure of video
3. Learn summarization by fooling a discriminator when trying to discriminate a machine-generated from a human-generated summary

Unsupervised approaches entails:

1. Learn summarization by fooling a discriminator when trying to discriminate the original video from a summary-based reconstruction of it
2. Learn summarization by targeting specific desired properties for summary
3. Build object-oriented summaries by modeling key motion of important visual objects

SUBCLASSES OF DEEP-LEARNING-BASED VIDEO SUMMARIZATION APPROACHES



1. Generating ground-truth human-labeled training data is highly expensive and time consuming
2. A video can have multiple summaries based on annotations from multiple human annotators
3. Ground-truth summaries might be very different from each other making it hard to train using typical supervised training methods

POWER OF UNSUPERVISED METHODS OF SUMMARIZATION:

WHY SHOULD IT BE PREFERED?

EVALUATION PROTOCOLS AND MEASURES FOR VIDEO STORYBOARDS

- Evaluation of key-frame-based summaries using human judgement
- Evaluation of generated summary according to its overlap with predefined key-frame-based user summaries
- Precision score
- Recall score
- F-Score

EVALUATION PROTOCOLS AND MEASURES FOR VIDEO SKIMS

- Evaluation of video skims according to their alignment with human preferences
- F-score

FUTURE SCOPE

- Research work and developments in the field of unsupervised video summarization
- Development of multimodal summarization approaches
- Advanced multi-head attention mechanisms
- Extension of LSTM architectures with high-capacity memory networks
- Combinations of architectures that use both- 3D-CNNs and convolutional LSTMs
- Use of augmented training data in combination with curriculum learning approaches
- Development of better evaluation measures for accurate performance comparison of different summarization methods.

CONCLUSION

- Upon analyzing the best performing models in **supervised video summarization**, it has been observed that they learn frames' importance by modeling the variable-range temporal dependency among video frames/fragments with the help of Recurrent Neural Networks and tailored attention mechanisms.
- These models have shown even better performance using extension of the memorization capacity of LSTMs by using their memory networks.
- Upon analyzing the best performing models in **unsupervised video summarization**, using Generative Adversarial Networks for learning how to generate a representative video summary could show promising outputs making them a forerunner and hence, this area needs to be studied further.
- For videos that do not possess any pattern and are very different from each other, GANs could work really well. These unsupervised techniques could be just the start of a new era in deep learning technology when it comes to video summarization!