Video Summarization

Sagar
Computer Science
IIIT Sri City

Pawan Kalyan Dakkata
Computer Science
IIIT Sri City

Wasim Ishaq Khan
Computer Science
IIIT Sri City

Shubham Gupta
Computer Science
IIIT Sri City

Pradeep Turan
Computer Science
IIIT Sri City

***Abstract*— In this project we propose a novel method for supervised, keyshots based video summarization. We propose a deep learning based architecture to perform video summarization based on emotions described by the speaker.**

Keywords—video summarization, face detection, emotion detection, frame extraction.

I. Introduction (*Heading 1*)

Conference videos and videos in many other domains are becoming to dominate other forms of information exchange. According to Cisco Visual Networking Index: Forecast and Methodology, 2016-2021 3, by 2019 video will account for 80% of all global Internet traffic, excluding P2P channels. Consequently, better methods for video management, such as video summarization, are needed.

An ideal video summarization is that can provide users the maximum information of the target video with the shortest time. Its goal is to produce a compact yet comprehensive summary to enable an efficient browsing experience. The video summary need to convey most of key information contained in the original video.

Video summarization is a task where a video sequence is reduced to a small number of still images called keyframes, sometimes called storyboard or thumbnail extraction, or a shorter video sequence composed of keyshots, also called video skim or dynamic summaries. The keyframes or keyshots need to convey most of key information contained in the original video. This task is similar to a lossy video compression, where the building block is a video frame. In this paper we focus solely on the keyshots based video summarization.

The architecture is centered around two key operations, emotion detection and frame level score. Frame score at every step t is estimated. Given the generic architecture of our model we believe that it could be successfully used in other domains requiring sequence to sequence transformation.

Our contributions are:

1. A novel approach to sequence to sequence transformation for video summarization based on face emotions. In contrast, the current state of the art relies on complex LSTM/GRU encoder-decoder methods.

2. A demonstration that a recurrent network can be successfully replaced with simpler, mechanism for the video summarization that shows the least neutral shots in the video which refers that the shot has a strong emotion to convey.

II. Our Approach

A. *Frame Extraction*

Video is broken down into frames and 1 frames is taken from every second.

B. *Face and Emotion Detection*

Faces are detected in all the extracted frames. Emotion of the extracted faces are then detected to determine the emotional score of every frame.

C. *Frame Ranking & Selection*

Frames having strong emotions are collected and then ranked.

D. *Subtitles Extraction*

Video Subtitles are prepared which represent the text along with the timestamps. From the subtitles, the sentence corresponding to these selected frames are extracted along with the starting and ending time of the sentence. Summary is prepared by merging video corresponding to all these intervals.

III. Frame Scores to Keyshot Summaries

A. *Frame Extraction*

In this process, we extract a single frame per second. Emotions of the extracted frames are detected to determine the emotional value of every frame.

B. *Emotion Detection*

- To detect the emotions, the extracted frames are passed on to the deep learning model.
- The deep learning model is implemented in two stages, one is Face Detection and the other is Emotion Classification.

- Face Detection is done using SSD (Single Shot Multibox Detector). The bounding box obtained from SSD is passed on to a CNN based emotion classifier.
- SSD speeds up the process by eliminating the need of the region proposal network. The SSD object detection composes of 2 parts: Extract features maps and apply convolutional filters to detect objects.
- SSD uses VGG16 to extract feature maps. Then it detects objects using the Conv4_3 layer. Each prediction composes of a boundary box and 21 scores for each class and we pick the highest score as the class for the bounded object
- The bounding box obtained from SSD is passed on to a CNN based architecture. This architecture of emotion classification is trained on FER2013 dataset. It presents the probability distribution of 7 different emotions as follows Disgust, Happiness, Fear, Sadness, Surprise and Neutrality.
- The testing accuracy of this classifier is 65% (SOTA is 71%).

C. *Frame Selection*

After the emotion score for each frame is estimated, the frames are then sorted based on the neutral emotion score. Then top k frames are selected (k depending on the desired length of the video).

If multiple faces are there in the frame, then for every face we are finding the emotion score and averaging it with the emotion score of the rest of them.

Frame (Average Probability)	F1	F2
Sad	0.23	0.301
Surprise	0.031	0.01
Angry	0.598	0.100
Happy	0.051	0.287
Fear	0.031	0.098
Disgust	0.073	0.016
Neutral	0.013	0.240

The table above shows two frames along with emotion probability distribution. F1 is selected, because neutral value has the less value compared to other emotions. And F2 is discarded since the surprise emotion has the lowest value.

Frame (Average Probability)	F1	F2
Sad	0.23	0.301
Surprise	0.031	0.240
Angry	0.598	0.109
Happy	0.051	0.287
Fear	0.031	0.098
Disgust	0.073	0.016
Neutral	0.013	0.01

From the above table, we can infer that the neutral value for both the frames are low. Since the value of emotion in the frame F1(Angry) is more than the frame F2 (Sad), the frame F1 will be hierarchically above F2 in the sequence.

The selected frames will be processed further.

D. *Subtitles Extraction*

- The subtitles are collected from youtube.
- The subtitles comprise of transcript along with corresponding timestamp.
- Given a time of a frame, corresponding point would be detected in the subtitles. Starting and ending point of the sentence would also be detected.
- Timestamp of the starting and ending of the sentence would be calculated. Video is extracted from these timings.

IV. Conclusion

We are able to summarize a speech type video in a continuous trailer like format without any gap or breakthroughs. The model is able to provide a space and information efficient summary.

References

[1] Using deep learning to find basketball highlights. Link: https://medium.com/in-the-hudl/using-deep-learning-to-find-basketball-highlights-edd5e7fa1278

[2] Video Summarization Using Deep Semantic Features by Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkila, Naokazu Yokoya.

[3] Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2015.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in ECCV, 2016.