Multi-Modal Video Summerization

Shubham S Patil
shubhamp@andrew.cmu.edu

Abstract

Ever growing consumption of online videos from search, recommendation, and sharing has generated a strong demand on compact summarization, to allow users to quickly understand the video content either as image thumbnails, and brief textual description from audio-visual features. This project aims to fuse the visual and auditory modalitites for summerization through event detection.

1 Project Goal & Impact

The goal of the project is to summarize a video, containing audio in a compact form fusing both the modalities. Given the immense video content available on the internet, effective summerization would help in recommendation systems, surveillance, event detection, etc. The main motivation of the project are:

1. No single modality is perfect in summarizing a video. Some videos need more emphasis on the visual aspect while others require auditory attention. Some might require both.
2. Rejecting either video or audio (i.e. modality) risks discarding valuable information.
3. Manually summarizing a video is expensive and time consuming. Given the immense data over the internet, it is practically impossible to provide summary of every possible video out there.
4. Event detection can be both audiovisual.

2 Literature Review

Summaries of video data may be static or dynamic; personalized, domain-dependent or generic; interactive or unsupervised; based on objects, events or perceptual features, such as user attention. Attempts to incorporate multi-modal and/or perceptual features have led to the design and implementation of various systems that take into account more than the visual stream of a video. Top three relevant papers to this project are:

1. Summarization based on Aural, Visual,and Textual Attention
2. Audio salient event detection and summerization using audio and text modalities
3. Compact Web Video Summarization Via Supervised Learning

Video summerization using DNN [3] : In this paper, instead of manually picking visual saliency features, the paper focuses on feature learning power of Convolution Neural network. This paper transforms the video summarization task into a image similarity learning task. [2] describes a synergistic way of combining textual and audio features for robust summerization.

Summerization based on Aural, Visual, Textual attention [1] discuss in detail of how all various modalities can be extracted and fused linearly or non-linearly. Although it produces great results, largely the features were handcrafted, data manually annotated, and didn't use any state of the art implementation using Deep Learning.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

3 Implementation and Methodology

In this project, we will be looking at the individual summerization as well as the performance improvement by using fusion. While working with any modality, we will be extracting the salient features. The entire project can be subdivided into the following tasks:

3.1 Audio Summerization

We will be looking at MFCC features, same as we did in assignment 1. However, We take a synergistic approach to audio summarization where saliency computation of audio streams is assisted by using the text modality as well. Inspiration has been drawn from [2].

3.2 Visual Summerization

The summarization of video's visual features is a highly subjective task and it varies a lot across annotators. Therefore, it is difficult and inappropriate to manually predefined any criteria or features of videos to model the summarization decisions of the human. Instead of handpicking features, we train a Deep Neural Network to learn the features and criteria used by human for video summarizationfrom labeled videos. The loss function used in [3] will be used to train a DNN model over the training set. The output of the summerization will be thumbnails.

3.3 Textual Summerization

I will be using Automatic Speech Recognition (ASR) to convert audio into text and summerizing them. This will be mostly through IBM or Google Speech-to-text API. This will be part of the language understanding. Additionally, Audio summerization is know to benefit from semantic understanding and hence, this will help with audio summerization as well.

3.4 Multi-Modal Fusion

There are two important fusion modalities namely Intradmodal and Intermodal.

1. Intra-modal fusion: Features for each modality are normalized and combined to produce modality-specific saliency modalities.
2. Inter-modal fusion: Saliency features from different modalities are combined in acomposite, multimodal saliency

In this project, we will be looking at Inter-modal fusion with Linear Fusion and Variance based fusion based on the features extracted from visual, auditory, and textual corpus.

4 Dataset

For this project, VSUMM (Video SUMMmerization) [4] and Carnegie Mellon University Movie Summary Corpus [5] datasets will be used.

VSUMM data-set videos are in MPEG-1 format (30 fps, 352 x 240 pixels), in color and with sound. These videos are distributed among several genres (documentary, educational, ephemeral, historical, lecture) and their duration varies from 1 to 4 minutes and approximately 75 minutes of video in total. Each video has 5 visual summaries

The CMU Movie Summary Corpus [5] contains 42,306 movie plot summaries. These summaries, contain a concise synopsis of the movie's events,along with implicit descriptions of the characters.

5 Metrices for Success

We will be compairing the results against VSUMM data-set [4] for visual summerization alone. The data-set also hosts results from various fusion algorithms. For Audio-Semantic summerization, we will compare it against the CMU benchmark in [5]. For Fusion, the results will be compared against those stated in [1].

The final report will have thumbnails of major visual events along with textual summary of the video.

6 Esitmated Computation Resources

I am planning on using AWS EC2 for training for neural network for visual summerization. I might use AWS for summerizing large videos and for storage purpose. I am expecting computation resource to cost around $50.

7 Timeline

Following is the timeline for the project

	March 26	**Project Draft**
	April 01	Extract MFCC Features, Speech-to-text, and fuse the Audio-semantic results
	April 05	Priliminary results of Visual summerization Network Training
	April 06	**Midterm Presentation**
1	April 12	Complete training Visual Summerization Network
	April 16	Fuse Visual Summary with Audio summary
	April 20	Try Various Fusion Models listed in [1]
	April 27	**Final Presentation**
	May 06	**Final Report**

References

[1] G. Evangelopoulos et al., "Video event detection and summarization using audio, visual and text saliency," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, 2009, pp. 3553-3556.

[2] Athanasia Zlatintsi, Elias Iosif, Petros Maragos (2015) Audio salient event detection and summerization using audio and text modalities. *23rd European Signal Processing Conference*, , Nice, 2015, pp. 2311-2315.

[3] Y. Wang, B. Han, D. Li and K. Thambiratnam, "Compact Web Video Summarization Via Supervised Learning," 2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW), San Diego, CA, 2018, pp. 1-4.

[4] Sandra E. F. de Avila, Ana P. B. Lopes, Antonio da Luz Jr., Arnaldo de A. Araújo VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method Pattern Recognition Letters, Volume 32, Issue 1, January 2011, pages 56–68

[5] Learning Latent Personas of Film Characters David Bamman, Brendan O'Connor, and Noah A. Smith ACL 2013, Sofia, Bulgaria, August 2013