Script Retrieval from Video Subtitle Data

Mohit Gupta[1]
Advisor: Dr. Amitabha Mukerjee[2]
{mohitt, amit}@iitk.ac.in
[1] *Dept. Of Electrical Engineering, IIT Kanpur*
[2] *Dept. Of Computer Science and Engineering, IIT Kanpur*

October 21, 2013

1 Introduction

As the size of multimedia archives continue to soar, its management by analysis and classification becomes a crucial task. By tapping into this data, we can unlock a huge realm of possibilities.

Which movie it was which had a scene where *"a black guy and a white guy were talking about some guy giving someone's wife a foot massage"?* It is reasonable to assume that it is a question only the people who have seen the movie will be able to answer and a computer would not be considered 'intelligent' enough. It is largely because humans store not an incessant stream of images and audio but a rather high level abstraction as a sequence of events and actions that take place or the changes in the state of the movie domain. For e.g. when Batman says "I am Batman", we do not record the audio signal, but we record what was said and can recreate it in our memory using Batman's voice signature.

If a computer could condense multimedia information like this, it could allow new efficient multimedia storage techniques. By learning from a semantic representations of a movies, a computer could then be trained to produce a movie, consuming a ridiculously less amount of time and resources. Summarization or filler removal by separating the mainstream storyline from the superfluous fragments would be another open possibility.

2 Related Work

Of the different streams in a video visual, audio and subtitles, subtitles provide quite a direct access to the semantic aspect. There have been not many attempts towards using NLP techniques on subtitle data. In [1], an approach for semantic video indexing and keyword extraction from subtitle data is described. Video categorization by keyword extraction from subtitles is another such example demonstrated in [2].

3 Approach

Before proceeding for semantic acquisition from subtitles, a crucial and challenging task is to get the script, i.e. classify different scenes, different speakers and annotate who says what.

I propose an approach for retrieving the script from a video by utilizing its subtitles, audio and visual content. Speakers can be separated by face recognition or voice recognition approach, of which voice recognition seems to be the viable option as it may not always be the case that the speaker's

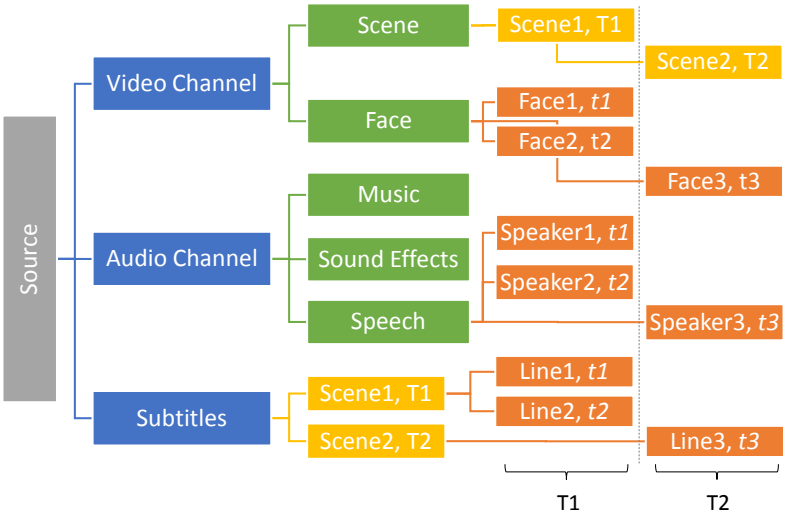

Figure 1: Flowchart showing the temporal correspondences of lines, speakers and faces, and their partitioning into scenes

face be shown while he speaks.

Annotating the audio for when someone is speaking and when some soundtrack or other sound effects are playing is the foremost task. Next, it is required to cluster the different voices and the corresponding lines in the subtitle files. It can be based on extraction of MFCC (mel-frequency cepstrum coefficients), which has been a successfully used feature for audio classification. After clustering the speakers, annotate a face to each from the video would be the final step of speaker identification.

For scene partitioning, a thresholding approach to the temporal gap between elements in the subtitle file can be used. The frame similarity in the visual stream or black frames could be a clue. Also, the change in the set of speakers relative to the previous scene could be another heuristic. The approach has been illustrated as a flowchart in Fig. 1.

4 Dataset

The dataset I plan to test this approach on will be a few episodes from Friends (TV series).

References

[1] Yi, Haoran, Deepu Rajan, and Liang-Tien Chia. "Semantic video indexing and summarization using subtitles." Advances in Multimedia Information Processing-PCM 2004. Springer Berlin Heidelberg, 2005. 634-641.

[2] Katsiouli, Polyxeni, Vassileios Tsetsos, and Stathes Hadjiefthymiades. "Semantic Video Classification Based on Subtitles and Domain Terminologies." KAMC. 2007.

[3] Shah, Sejal, and Archana Bhise. "Fast Speaker Recognition using Efficient Feature Extraction Technique." International Journal of Computer Science 2.