Abstract

Video segmentation is the first and critical step in video indexing and retrieval. Previous work in this area has primarily focused on visual and audio information. In this paper, we investigate the segmentation of documentary video data through audio and text understanding. To segment a continuous documentary video stream into subtopic segments, music markers and domain-independent video speech text segmentation are explored. Experiments are presented and discussed.

1. Introduction

Advances in computer power, network bandwidth, information storage, and signal processing techniques have led to a proliferation of video data. To support flexible video indexing and retrieval, the first and critical step is to segment a continuous video stream into a set of semantically tractable units.

Video segmentation has been studied for many years. Earlier works in this area have primarily focused on using visual information. Recently, researchers have begun to realize the importance of audio information [1, 2, 3, 4, 5, 6, 7, 8]. But most of these works make use of either audio features, such as volume and pitch, or video transcript. In this paper, we are interested in investigating the semantic segmentation of documentary videos through both audio and text analysis.

Two levels of audio features have been studied in audio content analysis: short-term frame level and long-term clip level [5]. An audio frame is defined as a group of neighboring samples that last about 10 to 40 milliseconds, within which we can assume that the audio signal is stationary. Frequently cited frame-level audio features include volume, Zero Crossing Rate (ZCR), pitch, frequency centroid [9], Bandwidth [9], spectral rolloff point [10], and MFCC [11]. To study the semantic content of a video, we often need to observe clip-level audio features. An audio clip is defined as a group of neighboring samples that last about one second to several tens seconds. Four types of clip-level audio features have been studied in the literature [5]: volume based, ZCR based, pitch based and frequency-based. Music marker, the audio feature we used in this paper, is built based on clip-level audio features. It is a higher level audio feature than both frame-level and clip-level audio features.

Video speech is a rich information source. Segmentation of transcribed video speech has been studied for some years. Work in this area has been largely motivated by the topic detection and tracking (TDT) initiative [6, 7, 8]. Most of the proposed approaches are suitable for broadcast and news domain in which the formal presentation format and cue phrases can be explored to improve segmentation accuracy, and there is a large set of training data available. In contrast to the TDT text segmentation study, several domain-independent text segmentation algorithms have also been proposed. Domain-independent text segmentation is derived from the theory of lexical cohesion [12, 13, 14], which states that text blocks with similar vocabulary are likely to be part of a coherent topic. Thus, finding topic boundaries could be achieved by detecting topic transitions that are indicated by vocabulary change. Most of the work in domain-independent segmentation focuses solely on text segmentation. In this paper, we are interested in investigating how domain-independent video speech segmentation aids semantic segmentation of documentary videos.

The rest of the paper is organized as follows. In section 2, we define what is meant by “subtopic segment” in this paper, which underlies the segmentation objective described later on. Section 3 describes how music markers segment a documentary video into semantic segments, each of which has a distinct subtopic, and Section 4 investigates how video speech aids the segmentation. Experimental results are presented in Section 5 and the paper concludes in Section 6.
2. What is subtopic segment?

In order to describe the process of dividing a continuous documentary video into subtopic segments, it is important to define what is meant by subtopic segment.

In our study, we view a documentary video as a long document with multiple subtopics. The major aim of our work is to investigate how effectively audio and text information can semantically segment documentary videos into subtopic segments. Although video segmentation work is often done at a finer granularity level than suggested here, subtopic segmentation has many potential applications, such as video indexing and retrieval. The ultimate goal of subtopic segmentation is to identify and label the subject matter. In a video retrieval scenario, given a query topic, a video retrieval system should only present video segments that are relevant. Thus, video segment boundaries should match topic boundaries as close as possible.

Defining the notion of topic is not an easy task. In discourse analysis, a topic generally means one stretch of discourse that ‘about’ something and the next stretch ‘about’ something else. Yet in most cases the basis for the identification of ‘topic’ is rarely explicit. Instead of pining the concept down, the notion of topic shift has been suggested and investigated [15].

That is, between two contiguous pieces of discourse which are intuitively considered to have two different ‘topics’, there should be a point at which the shift from one topic to the next is marked. If we can characterize this marking of topic shift, then we shall have found a structural basis for dividing up stretches of discourse into a series of smaller units, each on a separate topic……

The notion of subtopic segment adopts this stance, i.e., discovering topic shifts rather than identifying subtopics. Thus, the task of segmenting a documentary video into subtopic segments is converted to the process of discovering topic shifts.

A topic or subtopic shift in a documentary video stream is often associated with simultaneous changes of visual and audio characteristics. Since visual changes indicate shot breaks and video shots are often over-detailed for the extraction of high-level video semantic [16], audio information, including audio features and video speech, is generally more important in identifying topic shifts than video shots. In the following two sections, we will show in details how to detect subtopic shifts using this information.

3. Detecting subtopic shifts via music markers

The concept of music marker comes from the study of documentary video structures. In our study, we found that most documentary videos have background music. Filmmakers intentionally use music to structure the content and communicate the semantic. There is a temporally repetitive pattern of interleaving speech and music in videos. In general, videos start with a music segment that introduces the context and end with a music segment that summarizes the video content, predicts the future and sometimes presents copyright information. In the middle, speech and music segments of variable lengths alternate. In short,

\[\text{Documentary} \Rightarrow \text{Intro-music} \{\text{speech}\} \{\text{music}\} \Rightarrow \text{End-music}\]

This structure suggests that interleaving music segments could be used for subtopic detection. Based on the observation, we present a three-step procedure for subtopic detection, i.e., feature calculation, speech/music classification, and subtopic segment detection.

3.1. Feature selection

To use music in semantic video segmentation, we need to differentiate background music from speech. A wide variety of audio features for speech/music discrimination have been studied. Grounded on the work [7, 17], we select three groups of clip-level audio features: ZCR based, volume based, and spectral flux. These features describe the variations of ZCR, short time energy, and spectrum of an audio clip.

ZCR based: Zero Crossing Rate is the number of time-domain zero crossings within an audio frame. Three statistical measures of ZCR are used. They are i) standard deviation of the first order difference, ii) the third central moment about the mean, and iii) the difference between the number of audio frames whose ZCRs are above and below the mean value of an audio clip.

Volume based: volume is approximated by the Root Mean Square (RMS) of the signal magnitude within an audio frame. Three statistical measures of RMS are employed. They are i) standard deviation of the first order difference, ii) the third central moment about the mean, and iii) the difference between the number of audio frames whose RMSs are above and below the mean value of an audio clip.

Volume based: volume is approximated by the Root Mean Square (RMS) of the signal magnitude within an audio frame. Four statistical measures of RMS are employed. They are i) standard deviation of the first order difference, ii) the third central moment about the mean, and iii) the difference between the number of audio frames whose RMSs are above and below the mean value of an audio clip, and iv) low short-time energy ratio, which is the percentage of frames with RMS less than 50% of the mean value.

Spectral Flux: spectral flux indicates frame-to-frame spectral amplitude difference and is represented
using \(|\|X_i - X_{i+1}\|\). The sum of the differences of one audio clip is used in this paper.

After feature extraction, feature values are normalized using Gaussian Normalization [18] across both classes, i.e., speech and music, so that each feature has equal weight in the following subtopic detection processes.

### 3.2. Speech/music classification

Due to its simplicity and comparable accuracy [7], we use KNN to classify an audio stream into speech and music segments.

In our study, we found the probability to observe a single speech segment (\(\leq 1\)s) surrounded by music segments is very low, and vice versa. Based on this empirical observation, we perform a simple fine-tuning on classification results. For each spurious segment, i.e., a single speech segment surrounded by music segments or a single music segment surrounded by speech segments, we adjust and predict its class label to be the same as its surrounding segments. After this adjustment, all speech/music segments are at least 2 seconds long.

The last step is to collect neighboring segments of the same class to produce continuous, long speech/music segments. The result of this step is a collection of temporal intervals of interleaving speech and music segments, i.e.,

\[
\{[stm(i), etm(i)], [stm(i+1), etm(i+1)], \ldots, [stm(n), etm(n)], [stm(n+1), etm(n+1)]\}
\]

where \(st\) represents start time and \(et\) represents end time, subscript \(m(i), s(j)\) \((1 \leq i \leq n + 1, 1 \leq j \leq n)\) indicate \(i^{th}\) music segment and \(j^{th}\) speech segment respectively.

### 3.3. Subtopic segment detection

In our previous study [19], we observed that music segments are of difference semantic levels and the lengths of combined segments, i.e., speech segment plus the music segment that follows, correlate well with the semantic levels. Bigger segments are associated well with higher level semantic breaks, while smaller segments often indicate lower level semantic breaks. Since the aim of this study is to segment a continuous documentary video stream into subtopic segments, which is of higher-level semantic segmentation, we focus on bigger segments.

To detect subtopic shifts, we first calculate combined segment lengths. Then, we cluster these combined segments into two groups based on these lengths, significant and insignificant, using K-means algorithm. Finally, for each music segment \([stm(i), etm(i)]\) in the significant group, we define its music marker \(tm(i)\) as the midpoint of that temporal interval (Formula (1)).

These music markers segment a continuous video stream into subtopic segments.

\[
t_m(i) = \frac{stm(i) + (etm(i) - stm(i))}{2} \quad 1 \leq i \leq n \quad (1)
\]

### 4. Detecting subtopic shifts via speech text segmentation

Video speech is a rich information source. Documentary videos can be any subject and may come from different domains without adequate amount of training data. Thus, to detect subtopic shifts, semantic embedded in speech is explored in this section. One of domain-independent text segmentation methods is TextTiling.

TextTiling[17] is a technique for dividing long expository texts into multi-paragraph units that represent subtopics. Based on the vector space model and lexical cohesion theory, topic boundaries are located where similarity between neighboring blocks is low. In this paper, we apply TextTiling to video speech. Specially, the following steps are followed:

- morphological analysis, lexical score determination, and boundary identification. Morphological analysis determines the terms to be used in the following phases. Lexical score determination measures the similarity between text segments, each of which is represented by a vector (Formula (2)) [14].

\[
\text{score}(i, j) = \frac{\sum_t w_t, b_i w_t, b_j}{\sqrt{\sum_t w_t, b_i^2 \sum_t w_t, b_j^2}}, \quad (2)
\]

where \(b_i\) and \(b_j\) are vectors that represents two text segments, \(t\) ranges over all the registered terms of \(t_i\) and \(C_j\), \(w_{t,b_i}\) is the weight assigned to term \(t\) in vector \(b_i\), and \(w_{t,b_j}\) is the term weight assigned to term \(t\) in vector \(b_j\). Here, the weights on the terms are simply their frequency counts. These scores can be plotted, text segment number against score. Boundary identification is based on depth scores, which measures the distance from the peaks on both sides of a given text segment and corresponds to how strong the cues are for a subtopic shift.

With TextTiling, we obtain multiple text segments, each of which has a distinct topic. Text-based segmentation result is then combined with that of audio segmentation.
5. Experiments

5.1. Experiments set up

The experiments use nine videos from Open-video Project, which are "Exotic Terrain" "NASA 25th Anniversary Show" "Airline Safety and Economy, Report #265" "Lake Powell" "Energy Gas" "How Water Won the West" "Take Pride in America" "The Miracle of Water" and "The Colorado". Out of the nine videos, the first five are used for training while the remaining four are used for testing.

To detect subtopic shifts via music markers, 250 audio training samples, each of one second long, are prepared for each class, i.e. speech and music. For feature calculation, a sampling rate of 44.1 KHz is used. Fixed-length one-second audio clip is taken as basic unit for feature calculation and speech/music classification, which is further divided into 25ms non-overlapping audio frames. To detect topic shifts via speech text segmentation, transcripts from the Open-video Project are used.

F-score is adopted for performance evaluation. It is defined as \( F = \frac{2 \cdot P \cdot R}{P + R} \), where P is precision, defined as the ratio of the number of hits to the total number of detected segments, and R is recall, defined as the ratio of the number of hits to the number of actual segments. The higher the F-score is, the better the segmentation accuracy.

To claim a hit, the boundary of a detected text segment have to match what is manually determined and/or the corresponding music marker as defined in Formula (3) has to be within the temporal interval of the nearest music segment that is determined manually. Formally, let \( t_{m(i)} \) be any music marker, \([s_{m(i)}, e_{m(i)}]\) be the nearest music segment determined manually, then,

\[
\text{Hits} = \{ t_{m(i)} | t_{m(i)} \in [s_{m(i)}, e_{m(i)}], 1 \leq i \leq n + 1 \}
\]

5.2. Results and Discussions

The aims of these experiments are to evaluate how effectively music makers can detect subtopic shifts and how well speech text segmentation can help subtopic segment detection. Testing data consists of six video samples from four testing videos, all of which are approximately ten minutes long.

To evaluate the effectiveness of music makers in subtopic segment detection, the three steps described in Section 3 are performed. The results are shown in Table 1. Figure 1 illustrates one segmentation example (Note all tables and figures are at the end of the paper).

From Table 1, we can see that music markers can detect subtopic shifts reasonably well. However, not all subtopic shifts are indicated by music markers. Film makers do use other approaches to structure a video. In addition, if the background music has a lot of abrupt and striking changes, the algorithm often mistakes music for speech, which results in false subtopic shifts. Experiment No. 3 is such an example. Six extra segments have been falsely generated due to dramatic background music that the video has. This reasoning is justified by its low precision and high recall.

The detection effectiveness is also affected by songs. Unlike speech with background music, where the speech content dominates the music content, the singing content and the music content of a song are both emphasized to various degrees. If a video contains a song, some parts will be categorized as music, while other speech, depending on the style of the song. More salient features and procedures are needed to improve the performance.

To test how well speech segmentation can help subtopic segment detection, we applied TextTiling algorithm to transcribed video speech from the Open-video Project. Specially, the steps in Section 4 are followed. To integrate the results from speech text segmentation with those from music markers, we take the joint hits. A joint hit is defined as a hit from either music marker or speech text segmentation or both. In Table 2, "# OF DETECTED SUBTOPIC SEGMENTS" indicates the number of detected subtopic segments from either music markers or speech text segmentation or both. As it can be seen from Table 2, speech text considerably improves the performance of subtopic detection since the segmentation utilizes semantics embedded in video speech. The integration approach used here emphasizes the recall rates. If a higher precision rate is preferred, integration approach should be modified accordingly.

6. CONCLUSIONS AND FUTURE WORK

Video segmentation is the first and critical step in video indexing and retrieval. In this paper, we investigated documentary segmentation through audio and text understanding. We discovered that music markers can segment a continuous documentary video stream into subtopic segments reasonably well with an average F-score of 0.72 and speech text segmentation can considerably improve the segmentation performance with average F-score increased to 0.85. Using music markers for documentary video segmentation is a generic approach considering most documentaries use background music to help communicate. However, more salient features need to be studied for accurate speech/music classification with short time interval (≤ 1s), especially for documentaries...
with songs or striking music content. During text-based segmentation, parameters of the segmentation algorithm are manually tuned. In the future, automatic process should be studied using machine learning techniques.

References:


### Table 1: Subtopic Detection via Music Markers

<table>
<thead>
<tr>
<th>NO.</th>
<th>DURATION</th>
<th># OF SUBTOPIC SEGMENTS(&gt;2S)</th>
<th># OF DETECTED SUBTOPIC SEGMENTS (&gt; 2S)</th>
<th># OF HITS</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11 min 17 sec</td>
<td>16</td>
<td>15</td>
<td>12</td>
<td>0.8</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>2</td>
<td>9 min 59 sec</td>
<td>14</td>
<td>13</td>
<td>10</td>
<td>0.77</td>
<td>0.71</td>
<td>0.74</td>
</tr>
<tr>
<td>3</td>
<td>10 min 2 sec</td>
<td>9</td>
<td>15</td>
<td>9</td>
<td>0.60</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>11 min 32 sec</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>5</td>
<td>9 min 38 sec</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>0.63</td>
<td>0.71</td>
<td>0.67</td>
</tr>
<tr>
<td>6</td>
<td>9 min 40 sec</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>0.70</td>
<td>0.78</td>
<td>0.74</td>
</tr>
<tr>
<td>Ave.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.72</td>
</tr>
</tbody>
</table>
Figure 1. An Example of Subtopic Detection via Music Markers

Table 2: Subtopic Detection via Music Markers and Speech Text Segmentation

<table>
<thead>
<tr>
<th>NO.</th>
<th># OF SUBTOPIC SEGMENTS (&gt;2S)</th>
<th># OF DETECTED SUBTOPIC SEGMENTS</th>
<th># OF JOINT HITS</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>19</td>
<td>16</td>
<td>0.84</td>
<td>1.00</td>
<td>0.91</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>14</td>
<td>13</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>15</td>
<td>9</td>
<td>0.6</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>22</td>
<td>15</td>
<td>0.68</td>
<td>1.00</td>
<td>0.81</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>0.70</td>
<td>1.00</td>
<td>0.82</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>12</td>
<td>9</td>
<td>0.75</td>
<td>1.00</td>
<td>0.86</td>
</tr>
<tr>
<td>Ave.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.85</td>
</tr>
</tbody>
</table>

Figure 2: An Example of Subtopic Detection via Music Markers and Speech Text