Music Video Summarization

Hendra Setiawan and Edward Wijaya[1,2]
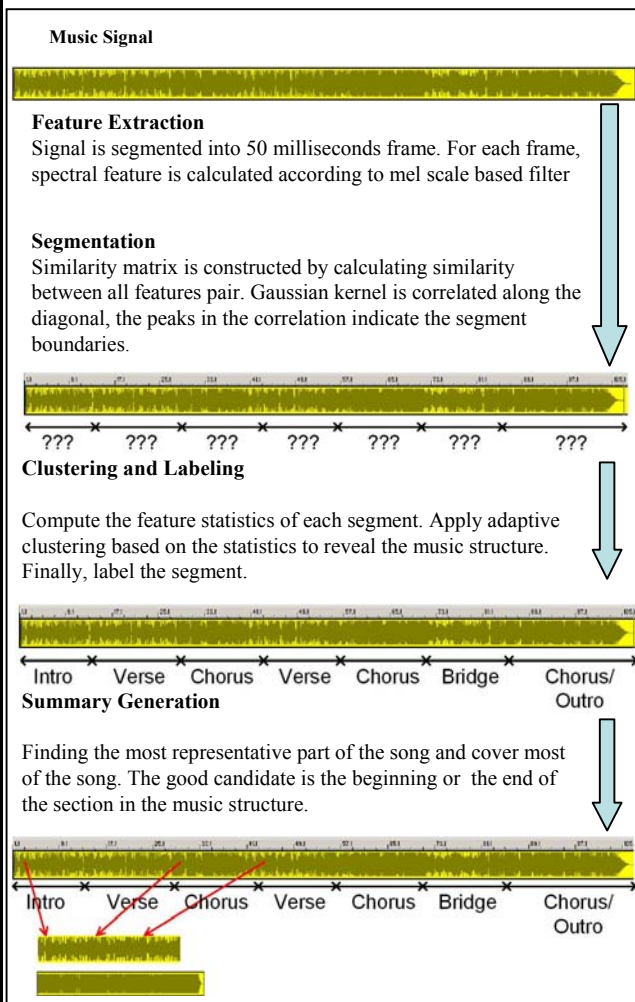
[1]School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543

[2]Institute for Infocomm Research (I2R), Heng Mui keng Terrace,
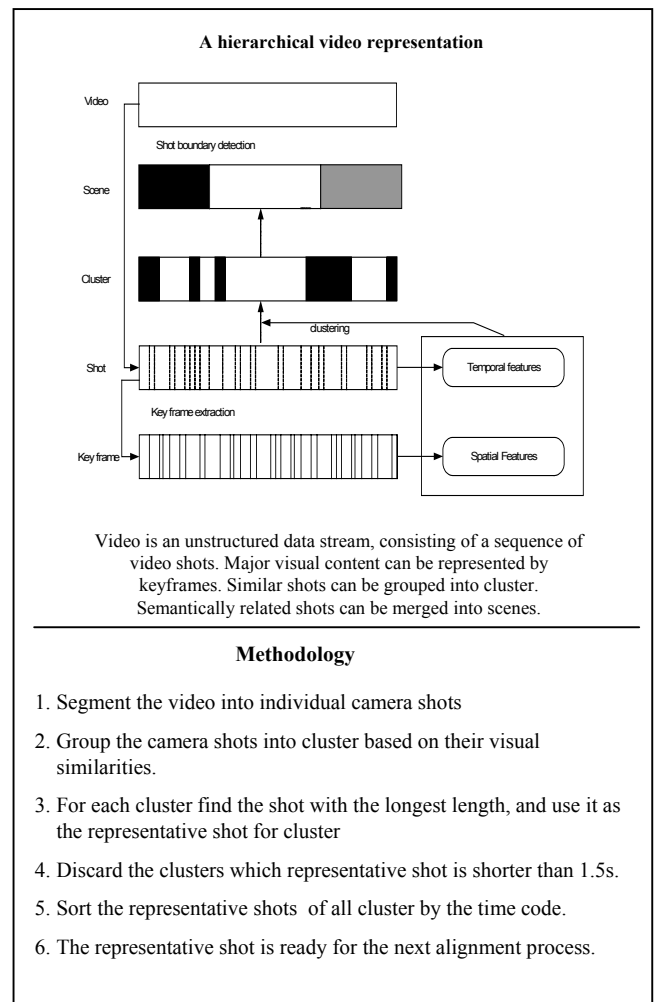Singapore 119613

Abstract and Motivation

We implemented music video summarization for our project. A good music video summarization techniques will enable user to quickly figure out a brief contents of the music video and help him to decide if it is worthwhile to watch the program. It enables the acceleration of information seeking process in multimedia digitial libraries. Our procedure, first separate the music video into audio and video track. Then we perform the summarization for each track. On video track summarization we do shot detection and clustering to obtain a smooth stream. Music summary is obtained through analyzing the music content, and carry out adaptive clustering with musical domain knowledge. Finally we align summary for each track to produce final summary of music video.

Music Summarization

Music Signal

Feature Extraction

Signal is segmented into 50 milliseconds frame. For each frame, spectral feature is calculated according to mel scale based filter

Segmentation

Similarity matrix is constructed by calculating similarity between all features pair. Gaussian kernel is correlated along the diagonal, the peaks in the correlation indicate the segment boundaries.

Clustering and Labeling

Compute the feature statistics of each segment. Apply adaptive clustering based on the statistics to reveal the music structure. Finally, label the segment.

Summary Generation

Finding the most representative part of the song and cover most of the song. The good candidate is the beginning or the end of the section in the music structure.

Video Summarization

A hierarchical video representation

Video is an unstructured data stream, consisting of a sequence of video shots. Major visual content can be represented by keyframes. Similar shots can be grouped into cluster. Semantically related shots can be merged into scenes.

Methodology

1. Segment the video into individual camera shots

2. Group the camera shots into cluster based on their visual similarities.

3. For each cluster find the shot with the longest length, and use it as the representative shot for cluster

4. Discard the clusters which representative shot is shorter than 1.5s.

5. Sort the representative shots of all cluster by the time code.

6. The representative shot is ready for the next alignment process.

MUSIC AND VIDEO ALIGNMENT

1. Summarize the music track of musical video. For example, the music summary which consist of 7 segments each which last 4.5 seconds, the total duration of summary is about 30 seconds.

2. Divide each music segment into three time slots each last for 1.5 seconds.

3. For each 1.5s time slot, the corresponding image is found as follow: we assign the image segment in time domain if it exist in representative shot set; if not find a most similar shot in the shot set. This can done using a similarity measure with feature like color histogram, shape, texture, and motion fields within frames in various scale. In our implementation we measure this similarity using color histogram.