Movie Indexing via Event Detection

Bart Lehane, Noel O'Connor

Centre for Digital Video Processing
Dublin City University
e-mail: {lehaneb, oconnorn}@eeng.dcu.ie

Abstract The past number of years has seen a large increase in the number of movies, and therefore movie databases, created. As movies are typically quite long, locating relevant clips in these databases is quite difficult unless a well defined index is in place. As movies are creatively made, creating automatic indexing algorithms is a challenging task. However, there are a number of underlying film grammar principles that are universally followed. By detecting and examining the use of these principles, it is possible to extract information about the occurrences of specific events in a movie. This work attempts to completely index a movie by detecting all of the relevant events. The event detection process involves examining the underlying structure of a movie and utilising audiovisual analysis techniques, supported by machine learning algorithms, to extract information based on this structure. This results in a summarised and indexed movie.

1 Introduction

Due to the huge amount of video in existence, management of digital video databases is problematic. Without explicit knowledge of the content, it is extremely difficult to locate relevant portions of a movie. As movies are usually quite long, this type of content is in particular need of indexing. However, as they are individually and creatively made, they are especially challenging to index. Manual annotation of video is both time consuming and expensive, so automatic indexing of content is clearly a desirable solution, but is also a challenging task. The aim of movie indexing is to facilitate efficient retrieval and browsing, and to ease the management of a movie database, so that a user can easily locate desired portions of a movie.

Much of the work in movie analysis has been in the detection of scene boundaries. For example, in [11,10], the authors use various measures of shot similarity to find points in a movie where one group of shots end and another begins. Although a scene-based structure is useful, it still contains a number of inherent problems. Firstly, in order to locate specific parts of the movie, all keyframes must be perused. Secondly, and most importantly, no semantic meaning is associated with the scene. To address these drawbacks, an event-based system is proposed that aims to classify each portion of the movie according to a relevant ontology. This aims to detect all relevant events in a movie, where an event is defined as a portion of a movie that moves the story onward. An event is a more fundamental unit than a scene, in that a scene may contain a number of different events e.g. a scene could start with shots showing two people talking (event 1), following which a fight breaks out (event 2), followed by more shots of people talking (event 3).

There have been a limited number of approaches to event detection in movies. In [6,7,2], the authors detect dialogues in video based on locating repeating shots of characters. The approach in [9] detects violent scenes in movies by locating flames, blood, explosions or screams. Although some event detection techniques achieve good results in isolation, very limited work has gone into detecting all relevant events, thereby building a complete summary of a movie. This work aims to index a movie by imposing an event-based structure on it. This involves selecting a number of event classes that cover all events in a movie, and detecting each event in each class.

2 Event Classification and Film Grammar

A number of event classes were created with which to index a movie. The event classes were created to be broad enough so that all of the relevant events in a movie belong to a class, yet meaningful enough so that it is possible for a user to class a new event into one of the existing event classes. The first event class consists of all *dialogue* events. This contains all conversations between characters in a movie. These conversations may involve any amount of people, in any setting (e.g a three-person conversation, a person addressing a crowd). The second event class contains all *exciting* events. This contains all portions of a movie in which excitement of the audience is desired (e.g. fights, battles, car-chases, arguments etc). The final event class contains a number of different event types, which, for convenience, are all labeled as *montage* events. The montage event class is a superset of three different events: montage events themselves, emotional events, and musical events. Although these events can differ greatly, a common characteristic is the presence of a strong musical background.

Movie making is an extremely creative process. Typically, film makers spend months writing, planning and shooting a single movie, which may result in less than two hours of video. Given this rather prolonged timescale, there is much room for creativity. Despite this, there are a number of film grammar rules that are universally followed in the world of movie making. These rules were initially created by the earliest directors, and have been refined by each new generation of film makers and still dictate many aspects of the way movies are filmed.

Generally, when trying to relax an audience, film makers use a relaxed shooting style. This usually involves using a static camera, in conjunction with a slow editing pace (i.e. quite long shots), a repetitive camera angle, and a relaxed audio track. There are many underlying reasons for this style. For example, a repetitive camera angle, where the camera remains in the same position for subsequent shots, ensures that an audience does not have to repeatedly interpret a fresh background. Similarly, a slow editing pace has the same effect. This allows the audience to focus on the activities on screen in a relaxed manner, without any distractions [1].

Conversely, when aiming to excite viewers, a faster editing pace, combined with camera movement, and non-repeating camera placement is used. This has the effect of bombarding the viewer with more visual information than it is possible to comprehend. As viewers aim to follow the activities as much as possible, an increased level of stimulation, and therefore excitement, is generated [3]. In general, the music in a movie is used to create an emotional response in the audience. Many film studios have libraries of music categorised by emotion, so when a film maker is looking to shoot a romantic scene, he/she goes to the 'romantic' song library. As [1] notes, sound effects are usually central to action sequences, while music usually dominates dance scenes, transitional sequences, montages, and emotion laden moments without dialogue.

3 System Overview

An overall system diagram for our approach is presented in figure 1. The input to the system is a movie file, while the output is an indexed movie in terms of the defined event classes. All of the analysis is undertaken using MPEG-1 video and PCM encoded WAV audio. The first step is the extraction of sub-shot features, corresponding to signal-based, low-level features. The features were chosen in order to garner as much semantic knowledge about the movie content as possible, which will in turn be used at a later stage. Thus, the features were chosen to detect the presence, and use, of the film creation tools mentioned in section 2.

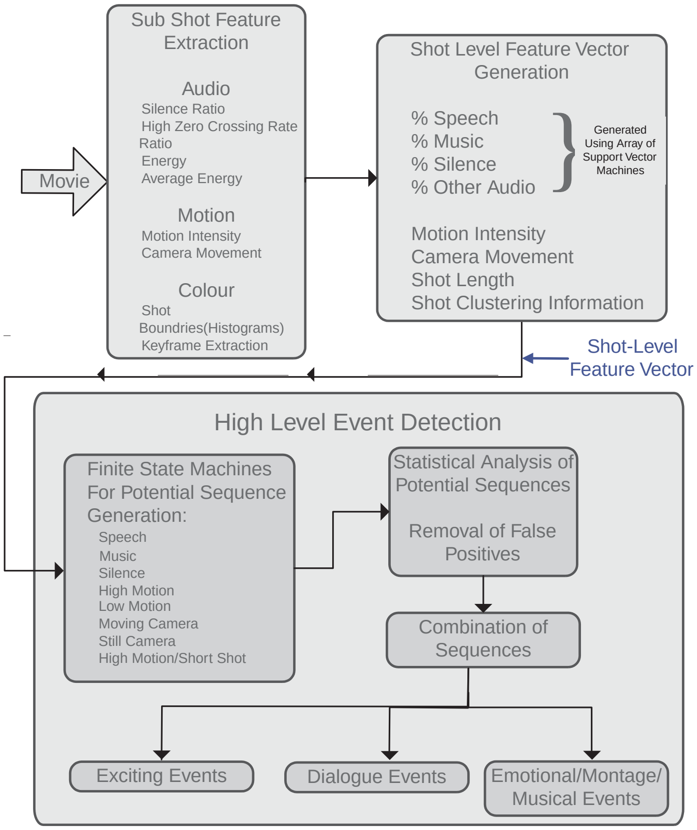
Movie
Sub Shot Feature Extraction
Audio
Silence Ratio
High Zero Crossing Rate Ratio
Energy
Average Energy
Motion
Motion Intensity
Camera Movement
Colour
Shot Boundries(Histograms)
Keyframe Extraction
Shot Level Feature Vector Generation
% Speech
% Music
% Silence
% Other Audio
Generated Using Array of Support Vector Machines
Motion Intensity
Camera Movement
Shot Length
Shot Clustering Information
Shot-Level Feature Vector
High Level Event Detection
Finite State Machines For Potential Sequence Generation:
Speech
Music
Silence
High Motion
Low Motion
Moving Camera
Still Camera
High Motion/Short Shot
Statistical Analysis of Potential Sequences
Removal of False Positives
Combination of Sequences
Exciting Events
Dialogue Events
Emotional/Montage/ Musical Events

Fig. 1 Overall System Structure

The sub-shot audio features were chosen in order to classify the audio into a number of relevant categories, which are useful for event detection. The features extracted are: silence ratio, high zero crossing rate ratio, short time energy, and average energy. Similarly, the motion features were chosen in order to accurately represent the amount and type of movement at any one time in the movie, as this can be used to infer knowledge about the film maker's intent. Thus, for each P-frame, camera motion is detected by analysing the MPEG-1 motion vectors, and the motion intensity value is calculated as the standard deviation of the motion vectors[8]. Colour histograms are also generated for each frame, and are then used in order to detect shot boundaries and select keyframes (one per shot).

Given the sub-shot features, a shot-level feature vector is then created. This involves taking the sub-shot features and creating a single value for each shot. For the audio, the sub-shot features are used as a basis for audio classification. Four audio classes are created: speech, music, silence and other (sound effects). These audio classes are based on observations about the type of audio content used in movies[1]. Using a support vector machine (SVM) based classification method, with the sub-shot audio features as inputs, the percentage of each of these audio classes per shot is calculated. So, for example, a single shot may contain 80% speech, 0% music, 20% silence and 0% other audio. A more detailed explanation of this process can be found in [4].

The two sub-shot motion values, extracted at each P-frame, are interpolated to produce a single value over the entire shot. Thus, each shot has a value for motion intensity, as well as a value for the amount of camera movement in the shot. Although there is a large amount of overlap between the two motion features, it is desirable to detect movement within the frame (motion intensity) as well as camera movement. There may be times where the camera is still, yet there is high amounts of activity on screen, or vice vesa.

Using the shot boundary information, the shot length can be calculated. This gives an indication of the speed of editing. Using the histogram of the keyframe for the shot, shots with similar colour are clustered into shots filmed from the same camera position. So, for example, in a conversation between three people colour shot clustering will result in three clusters corresponding to person A, B and C respectively. This is then used to detect changes of focus in the movie in a manner similar to the method proposed in[11].

The shot feature vector contains all of the audio, motion, and colour information for each shot. Thus, the complete shot-level feature vector contains the following values for each shot in a movie: [% speech, % music, % silence, % other audio, % static camera frames, % non-static camera frames, motion intensity, shot length]. This feature vector is then used for high level event detection. This process involves three steps. Firstly, an array of finite state machines (FSM) are used to create a set of *potential sequences*, these are areas in the movie where particular features dominate. For example, there is a speech FSM that detects all of the areas where speech shots occur frequently, resulting in a set of *speech potential sequences*. The second step involves filtering of the potential sequences so that they are either rejected, or labelled as *events* and placed into one of the event classes. Note that each event detection process is run independently. Finally, the events are combined to create a final event list, containing all dialogue, exciting, and montage events. The details of how each event is detected are presented in the following sections.

3.1 Detection of Dialogue Events

Characteristics of a dialogue event are clearly audible speech, a relaxed editing pace, a static camera that remains focused on the characters, and a high amount of shot repetition. In order to detect dialogue events, two FSMs are used: the speech FSM and the static camera FSM, since a dialogue event should contain at least one of these features. These two FSMs detect all of the areas with high amounts of speech (speech potential sequences), and static camera shots (static camera potential sequences). In order for a speech potential sequence to be classed as a dialogue event, it must contain either a high amount of shot repetition, or a high amount of static camera shots. In order for a static camera potential sequence to be classed as a dialogue event, it must contain either a high amount of shot repetition, or a high amount of speech shots. Both sets of events (i.e. from both the speech and static camera potential sequences) are then combined to produce one final list of dialogue events.

3.2 Detection of Exciting Events

Due to the reliance of filmmakers on using fast editing as well as high amounts of motion when filming exciting events, a combined FSM was created. This FSM detects all of the areas where high motion shots and short shots occur together. This results in a list of “high motion, short shot” potential sequences. These potential sequences are filtered to remove short, insignificant, bursts of excitement (less than five shots), as well as sequences that contain a lot of repetitive camera shots, as these are generally not present in exciting events[5].

3.3 Detection of Montage Events

Typically all of the events in the montage class contain a musical background. However there may be occasions where silence is used instead of music. Thus, the non-speech FSM is used (as opposed to the music FSM) when detecting these events. Once the non-speech potential sequences are generated, a number of filtering steps are undertaken. Montage events do not contain high amounts of shot repetition, therefore the non-speech potential sequences that contain very high amounts of shot repetition are removed. Also, if the non-speech potential sequences contain shots with high amounts of moving camera and fast-pased editing, then they are re-classed as exciting events (as many exciting events also contain a musical background), and merged with the previously generated exciting events. Following the detection of the three event types, the indexing process is complete.

4 Experimentation and Results

Ten movies were chosen as a test set, and were manually annotated in order to evaluate the proposed approach. The start and end time of each dialogue, exciting and montage event was noted by the authors. The

Table 1