Finite State Machines in Movie Scene Classification

Yun Zhai
University of Central Florida
Orlando, Florida, US
yzhai@cs.ucf.edu

Zeeshan Rasheed
ObjectVideo
Raston, Virginia, US
zrasheed@objectvideo.com

Mubarak Shah
University of Central Florida
Orlando, Florida, US
shah@cs.ucf.edu

Abstract

In this paper, we address the problem of the scene classifications in feature films and propose a robust framework for the stated problem. The framework utilizes the structural information of the movie scenes rather than analyzing the global low level feature values. We propose and demonstrate that the Finite State Machines (FSM) are suitable for classifying the movie scenes into three categories: conversation, explosion/gunfire and suspense. Three major characteristics of motion pictures, motion, color *and* audio, *are used in our approach. The transitions of the FSMs are determined by the mid-level features of each shot in the scene. Our FSMs have been experimented on a large set of data with both positive and negative examples and produces impressive results.*

1. Introduction

With the tremendous growth of media documents every year, finding efficient methods to understand and manage this vast amount of information becomes more and more challenging and urgent. As an important element of the entire entertainment industry, semantic labelling of feature films has attracted increasing attention from many researchers in various fields. Applications for content-based video annotation and retrieval have been developed at all levels of the movie structure: shot level, scene level, and movie level. A shot is a sequence of images that preserve similar background settings. It is the basic element of a movie. A scene consists of a set of continuous shots that constitute a particular story. On the top level, a movie is composed of a series of related scenes. For the movie audience, shot level analysis is insufficient to understand the video content. On the contrary, searching and returning an entire target movie might be computational complex and time consuming. To reach a reasonable balance between the information sufficiency and the processing cost, the analysis at scene level is more relevant. Furthermore, to accurately and fully express the theme of the scene, motion, color, and audio are the three major properties needed to be analyzed.

Currently, several works have been done on the topic of conversational scene detection and analysis. In [6], shot length and visual dynamic were considered in the process of scene type analysis. In this approach, only the repetition of similar shots was employed. In [4], the author used the pattern of the dialog shot. Color and audio information was used in the system. However, neither of these approaches addressed the use and importance of motion features in the scene. In [2], the author detected conversational scenes from independent or dependent auditory streams. The method involved some mid-level audio feature detectors, e.g., a joint voice and speech detection application. The system only utilized the audio cues of the video. The visual information was not mentioned. In [3], the system exploited the global structural information of a scene. They built "*shot sinks*" to classify a scene into one of three scenarios, "two speaker dialog", "multi-speaker dialog", and "others". The overall structure was computed based on the low level visual features of the shots in the scene.

In this paper, we present a new method for detecting conversational scenes in feature films. A Finite State Machine (FSM) is developed and implemented to fulfill this task. All three major properties, *motion*, *color* and *audio*, are used to build the system. The transitions are determined based on the statistics of the features for each shot. The rest of this paper is organized as follows: Section 2 describes the features we used and the methods for extracting these features from the videos. Section 3 demonstrates the structure and the use of the Finite State Machine. Section 4 shows the experimental results. Lastly, Section 5 concludes our work.

2. Feature Extraction

In this section, we present the features that are used in our Finite State Machines. In movies, the scenes are composed in accordance with the conventional film grammars. For conversational scenes, they have the following patterns: visual smoothness, medium audio energy and mul-

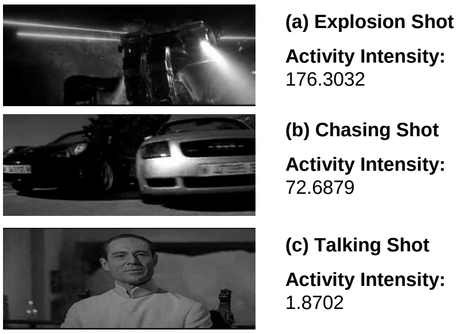


Figure 1. Three types of movie shots and their activity intensity values. (a) An explosion shot, (b) A car chasing shot, and (c) A talking shot. The left column are the key frames of the shots. Their activity intensity values are on the right.

tiple speakers. For each shot in the scene, we have two features. The *activity intensity* feature captures both the motion smoothness and the audio intensity. The *face identity* feature determines the membership of the speaker(s).

2.1. Activity Intensity (Γ)

In feature films, the camera motions are mainly either translation or zoom. Camera roll and tilt are rare. Therefore, we model the image-to-image global transformation between frames by an affine matrix. Instead of using information from every pixel, the motion field is computed based on the 16x16 macro-blocks. For each macro-block $[\ x\ \ y\ ]^T$, its motion $[\ u\ \ v\ ]^T$ is computed as

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & b_1 \\ a_3 & a_4 & b_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

The global translation is determined by vector $[\ b_1\ \ b_2\ ]^T$. The magnitude m of the translation vector represents the intensity of the motion, and the difference d between the directions of the motion vectors in adjacent frames gives the smoothness of the camera motion. Thus, an average scale $\lambda = (m + \kappa_m) \times (d + \kappa_d)$ over the entire shot can capture both the intensity and the smoothness of the global motion, where κ_m and κ_d are positive constants to avoid the multiplication of zero.

In conversation scenes, the local motion is not as significant as in other kinds of movie scenes. For example, in the scenes involving fighting or explosions, the local motion intensity becomes a major characteristic. We compute the local motion intensity as follows:

1. Estimate the affine motion parameters A from the given motion field U.

2. Apply the global motion parameters to the grid of the macro-blocks to establish a projected motion field U'.

3. For each frame in the shot, compute the mean difference between the original motion field U and the projected motion field U'.

4. Calculate the average μ of the mean differences for the entire shot.

Sound also plays an important role in the understanding of scenes. Unlike the scenes that involve explosions, collisions, or vehicle chases, the characters speak smoothly and calmly in conversational settings. For each shot in the scene, the mean audio energy θ is used.

Finally, we combine the three scales, λ, μ and θ, to form the activity intensity scale,

$$\Gamma = \lambda \times (\mu + 1) \times (\theta + 1) \tag{2}$$

Figure 1 shows the key frames and the activity intensity values of the examples for three types of shots: (a) explosion shot, (b) car chasing shot and (c) talking shot.

2.2. Face Identity

The presence of speakers constitutes the major part of conversational scenes. We use the knowledge that at least two speaking parties are needed to form a conversation. Because the speaking parties switch periodically, strong structure exists in conversations. Very commonly, these speaking parties can be found by their faces. Therefore, the scenes that contain few or no human faces are excluded. We detect the speaking parties by clustering their faces into different groups. The clustering process is described in the succeeding paragraphs.

The middle frame of each shot i is selected as the key frame k_i of that shot. The face detection program then is applied to the key frame k_i. Due to the various image sizes of different movies, the detection program is performed on four levels of scales to the original image size, 50%, 100%, 200% and 400%. The bounding boxes of the detected faces $F_i = \{f_i^1, ..., f_i^{n_i}\}$ are returned. For every face patch, a 24-bin RGB color histogram with 8-bin for each color channel is computed. The similarity $S(i,j)$ between two facial shots i and j is the similarity between the sets F_i and F_j,

$$S(i,j) = max(dist(f_i^m, f_j^n)) \tag{3}$$

where $m = 1, 2, ..., n_i$, $n = 1, 2, ..., n_j$, and $dist(f_i^m, f_j^n)$ is the histogram intersection between the image patches for the faces f_i^m and f_j^n.

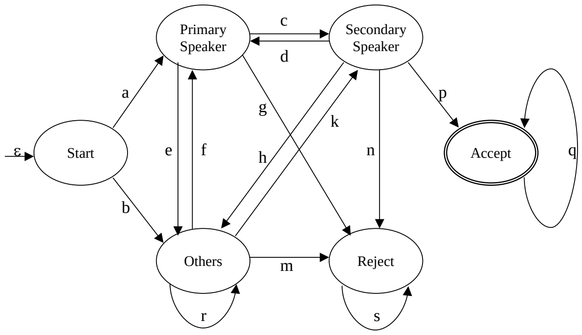


Figure 2. Finite State Machine for Conversation Scene Detection. It consists of six states.

The cluster is initialized by including the first facial shot to the empty set. For each new facial shot i, we find its membership with all the previous established clusters. The similarity $L(i,k)$ between the shot i and the cluster k is determined by:

$$L(i,k) = max(S(i,j)) \tag{4}$$

where j is the index of the shot in cluster k. The cluster that provides the maximum similarity score is declared as the group that the new shot belongs. However, if the score is below the threshold, a new cluster is constructed. This process is performed in real-time.

3. Finite State Machine (FSM)

We have built a deterministic Finite State Machine (FSM) (Figure 2) for detecting the conversation scenes. Our FSM consists of six states: “*Start*”, “*Primary Speaker*”, “*Secondary Speaker*”, “*Others*”, “*Reject*” and “*Accept*”. The state “Primary Speaker” is represented by the largest cluster, and the “Secondary Speaker” is represented by the second largest cluster. The transitions are determined based on the feature values of the shots in the scene. If the state is “*Accept*” at the end of the process, then the testing scene is declared as a “Conversation” scene. Otherwise, it is declared as a “Non-Conversation” scene.

The formal definition of the Finite State Machine (FSM) A is expressed by,

$$A = (Q, \Sigma, \sigma, q_0, F) \tag{5}$$

In this formulation, Q is the set of the states, where $Q = \{``Start", ``PrimarySpeaker", ``SecondarySpeaker", ``Others", ``Reject", ``Accept"\}$. q_0 is the set of starting states, and in our FSM, $q_0 = \{``Start"\}$. $F$ is the set of accepting states, and $F = \{``Accept"\}$. Σ is the range of the feature values.

The set of the transitions is denoted by σ. $\sigma = \{\varepsilon, a, b, c, d, e, f, g, h, k, m, n, p, q, r, s\}$. The transition matrix for σ is shown in Table 1.

σ	S	Pri	Sec	O	R	A
S	-	a	-	b	-	-
Pri	-	-	c	e	g	-
Sec	-	d	-	h	n	p
O	-	f	k	r	m	-
R	-	-	-	-	s	-
A	-	-	-	-	-	q

Table 1. Transition Matrix for σ. The column states are the “From” states, while the row states are the “To” states. Any entry that is not a “-” indicates a transition from one state to another.

The transition conditions are defined as follows:

- **“a”** - The first shot in the scene is a facial shot with low activity intensity. The transition transforms the state to “Primary Speaker”.
- **“b”** - The first shot in the scene is a non-facial shot with low activity intensity. The transition transforms the state to “Others”.
- **“d” and “f”** - The new shot is a facial shot with low activity intensity, and it belongs to the largest cluster. The transitions transform the state to “Primary Speaker”.
- **“c” and “k”** - The new shot is a facial shot with low activity intensity, and it belongs to the second largest cluster. The transitions transform the state to “Secondary Speaker”.
- **“e”, ‘h” and “r”** - The new shot is a non-facial shot with low activity intensity. Or, the new shot is a facial shot with low activity intensity but belongs to neither the largest cluster nor the second largest cluster. The transitions transform the state to “Others”.
- **“g”, ‘n” and “m”** - The new shot has high activity intensity. The transitions transform the state to “Reject”.
- **“p”** - The new shot is a facial shot with low activity key. It completes the accepting requirement of the FSM. The transition transforms the state to “Accept”.
- **“q”** - For any new shot, this transition loops at the state “Accept”.
- **“s”** - For any new shot, this transition loops at the state “Reject”.

4. Experimental Results

We have run the finite state machine on over 50 movie clips. These clips are selected from 7 movies, 2 Television talk shows (Larry King Live), and 2 Television news program (CNN Headline News). The movies cover almost all the genres of feature films, horror, drama, action, etc. The testing data set contains two portions: positive (conversation) clips and negative (non-conversation) clips. There are 20∼ 30 shots in each scene. To obtain the ground truth, four observers were asked to watch these clips and decide if they are conversational or non-conversational. Figure 3 shows four example testing clips with the key frames of the shots in the scene.

Two kinds of the accuracy scales were computed for each of the positive and negative testing sets: precision and recall. They are defined as,

$$P_{pos} = \frac{M_{pos}}{D_{pos}}, \;\; R_{pos} = \frac{M_{pos}}{G_{pos}} \qquad (6)$$

and

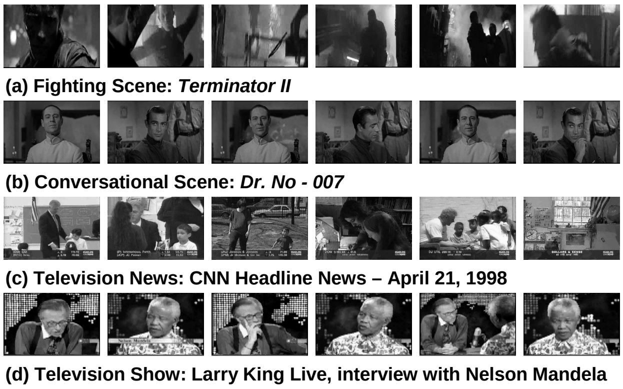
$$P_{neg} = \frac{M_{neg}}{D_{neg}}, \;\; R_{neg} = \frac{M_{neg}}{G_{neg}} \qquad (7)$$

where P_{pos}, R_{pos}, P_{neg} and R_{neg} are the precision and recall for positive and negative testing sets, respectively. G_{pos} and G_{neg} are the sizes of the ground truth positive set and the negative set. D_{pos} and D_{neg} are the sizes of the detected positive and negative sets. M_{pos} and M_{neg} are the numbers of the matched positive and negative movie clips.

There were 27 conversational scenes in the testing set. We achieved 96.2% precision and 92.6% recall. For the other 25 non-conversational scenes, the precision is 92.0%, and the recall is 95.9%. The accuracy scales demonstrate that the Finite State Machine can robustly detect the conversations from given movie scenes based on the proposed features. The scales are shown in Table 2.

5. Conclusions

In this paper, we have presented a new method for detecting conversational scenes in feature films by using a Deterministic Finite State Machine (FSM). The FSM consists of six states and the corresponding transitions between the states. We propose two features, the activity intensity and the face identity. The activity intensity provides the motion properties and the audio intensity of the shots in the scene. The face identity provides the speaker's membership. The transitions of FSM are determined by the instant feature values of the shots in the scene. We have tested the FSM on over 50 movie clips. The precision and recall for both positive and negative examples are high.



(a) Fighting Scene: *Terminator II*

(b) Conversational Scene: *Dr. No - 007*

(c) Television News: CNN Headline News – April 21, 1998

(d) Television Show: Larry King Live, interview with Nelson Mandela

Figure 3. Four Example Testing Clips. The first six key frames are displayed.

	Positive Set	Negative Set
Precision	96.2%	92.0%
Recall	92.6%	95.9%

Table 2. Precision and recall for both positive and negative testing sets.

Since this approach incorporates the temporal structural information of the video clips, it is reasonable and feasible to expand it for the detection of other types of movie scenes, which also present the temporal structures. We also experimented our methods on "Explosion" and "Suspense" scenes, and the results are promising.

References

[1] B. Adams, C. Dorai, S. Venkatesh, *Novel Approach to Determining Tempo and Dramatic Story Sections in Motion Pictures*, ICIP, 2000.

[2] S. Basu, *Conversational Scene Analysis*, Thesis, 2002

[3] Y. Li, S. Narayanan, C.-C. Jay Kuo, *Movie Content Analysis Indexing, and Skimming*, Kluwer Academic Publishers, *Video Mining*, Chapter 5, 2003.

[4] R. Lienhart, S. Pfeiffer, and W. Effelsberg, *Scene Determination Based on Video and Audio Features*, ICVIS, 1999.

[5] P. Viola, M. Jones, *Robust Real-Time Object Detection*, International Journal of Computer Vision, 2001.

[6] A. Yoshitaka, T. Ishii, M. Hirakawa, *Content-Based Retrieval of Video Data by the Grammar of Film*, IEEE Symposium on Visual Languages, 1997.