A Framework for Semantic Classification of Scenes Using Finite State Machines

Yun Zhai[1], Zeeshan Rasheed[2], and Mubarak Shah[1]

[1] School of Computer Science, University of Central Florida
Orlando, Florida 32828, United States
{yzhai,shah}@cs.ucf.edu
[2] ObjectVideo
11600 Sunrise Valley Dr. Suite 290
Reston, Virginia 20191, United States
zrasheed@objectvideo.com

Abstract. We address the problem of classifying scenes from feature films into semantic categories and propose a robust framework for this problem. We propose that the Finite State Machines (FSM) are suitable for detecting and classifying scenes and demonstrate their usage for three types of movie scenes; conversation, suspense and action. Our framework utilizes the structural information of the scenes together with the low and mid-level features. Low level features of video including *motion* and *audio energy* and a mid-level feature, face detection, are used in our approach. The transitions of the FSMs are determined by the features of each shot in the scene. Our FSMs have been experimented on over 60 clips and convincing results have been achieved.

1 Introduction

Recent years have seen a growing interest in the annotation and retrieval of video data. The increasing number of subscribers to digital cable now demands efficient tools so that viewers can browse and search sections of interest of video. Among many genres of video production, feature films are a vital field for the application of such tools. It is a sizeable element of the entertainment industry, easily available, widely watched and therefore, is becoming a focus of researchers in many aspects. For example, applications for content-based video annotation and retrieval have been developed at all levels of the video structure; shot level, scene level, and movie level. A shot is a sequence of images that preserve consistent background settings. It is the basic element of a movie. A scene, which consists of a set of continuous shots, constitutes a portion of the story line. On the highest level, a movie is composed of a series of related scenes defining a theme. For a user, who may be looking for a particular scene of a feature film, a shot level analysis is insufficient since a shot level analysis fails to capture the semantics of the video content. For example, how does one answer a query for a *suspense* scene in a feature film based on a single shot content? Any semantic category like *suspense* or *tragedy*, cannot be defined over a single shot. These concepts are induced in viewers over time. Indeed, a meaningful result can only be achieved by exploiting the interconnections of shot content.

P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp. 279–288, 2004.

In this paper, we present a novel framework for classifying scenes, focusing on feature films, into three semantic categories; conversation, suspense and action. This method analyzes the structural information of the scenes based on the low-level and mid-level shot features which are robust and easily computable. The low-level features used in our framework include shot motion and audio energy and the mid-level feature is face identity. To bridge the gap between the low and mid-level features and a high-level semantic category, Finite State Machines are studied and developed. The transitions are determined based on the statistics of these features for each shot. This paper is organized as follows: Related work is discussed in Section 2, Section 3 describes the classification framework, including the features and the Finite State Machines for detecting conversation, suspense and action scenes. Section 4 shows the experimental results and Section 5 concludes our work.

2 Related Work

In the area of higher level scene understanding, Adams et al [1] proposed the detection of "tempo" in movies. The camera motion magnitude and the shot length were the two features used to compute a continuous function. Our framework, on the other hand, analyzes the structure of the movie scene and classify scenes into more specific categories. Yoshitaka et al. [3] also used shot length and visual dynamics to analyze scene type. In their approach, the color statistics of the frames in the shot were used to calculate the visual dynamics and the similarities between the repeating shots were exploited. Experiments on only one kind of scene was demonstrated and it was not clear how the approach could be extended to other scene categories.

Lienhart et al. [4] used face detection in the scenes to link similar shots. A "face-based class" with a group of related frames showing the same actor was constructed by the similarity of the spatial positions and sizes of the detected faces. These "face-based classes" were linked across shots in the video to form the "face-based sets" by using Eigenfaces. The pattern of a dialog scene was flagged if several conditions were satisfied. In their experiment, face recognition suffered accuracy and the system typically split the same actor into different sets causing overdetection. Li et al. [5], exploited the global structural information of a scene and built "shot sinks" to classify a scene into one of three scenarios including *two speaker dialog*, *multi-speaker dialog*, and *others*. The overall structure was computed based on the low-level visual features, such as color of the shots in the scene. In their approach face information, which is an important cue for speaker detection, was not used. We combine both structure and face detection in a Finite State Machine framework to provide a more general solution for the scene classification task.

3 Proposed Approach

In this section, we first discuss the low-level and mid-level features used in our approach. The *activity intensity*, which is a function of low-level features and includes local and global motion and audio signal, is the input to the Finite State Machines. Human faces are detected in shots, clustered, and also used as input. We construct FSMs for three different semantic categories of scenes. These include conversational, suspense and action.

3.1 Computing Activity Intensity (Γ) Using Motion and Audio

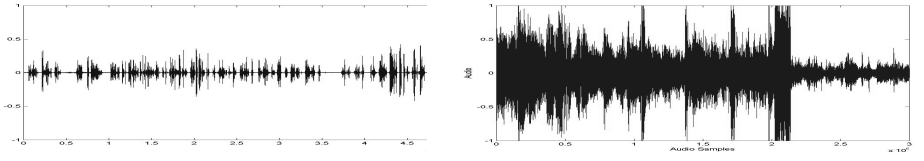
Motion in the videos has been used by several researchers in detecting and identifying scenes in feature films. Some examples are [1,2]. In feature films, the camera motion is generally translation and zoom, whereas, camera roll and tilt are rare. Affine motion model is suitable for capturing translation, scale and rotation about the optical axis of camera. Therefore, we model the image-to-image global transformation using an affine motion model. We exploit the motion vector information embedded in the MPEG compressed video. The approximate motion model is computed based on the 16x16 pixel macro-blocks. For each macro-block $[\ x\ \ y\ ]^T$, its motion $[\ u\ \ v\ ]^T$ is computed as

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

where $[\ b_1\ \ b_2\ ]^T$ vector captures the global translation. The magnitude m of the translation vector represents the intensity of the motion, and its absolute difference d across adjacent frames gives the smoothness of the camera motion. Thus, an average global motion quantity, λ over the entire shot captures both intensity and the smoothness of the global motion, that is:

$$\lambda = (m + \kappa_m) \times (d + \kappa_d), \tag{2}$$

where κ_m and κ_d are small positive constants to avoid multiplication with zero. The local motion intensity also varies with the type of scene. For example, in a fighting scene, high action intensity is commonly observed. We compute the local motion intensity by computing the mean difference, μ, of the reprojected motion vectors and the original motion vectors for the entire shot.

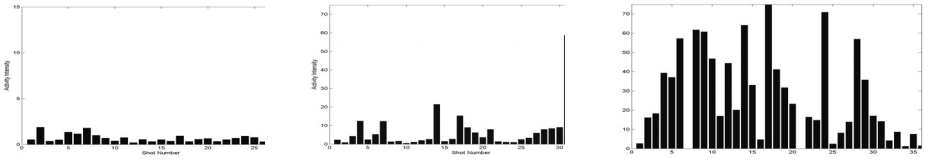


(a) (b)

Fig. 1. The audio signal for (a) conversation, and (b) action.

Sound also plays an important role in distinguishing scenes from each other. In conversational scenes, characters speak smoothly and calmly. In action scenes, which often include explosions, collisions, or vehicle chases, the audio energy is very high. Figure 1 shows the plot of audio signals for (a) conversational and (b) action scenes. Note that the high energy in the audio of the action scene is distinctive from that of the conversational scene. Therefore, the computation of activity intensity also incorporates the mean audio energy θ. The overall activity intensity is the combination of the three quantities, λ, μ and θ as follows:

$$\Gamma = \lambda \times (\mu + 1) \times (\theta + 1) \tag{3}$$



(a) (b) (c)

Fig. 2. The activity λ of three types of movie scenes (a) conversation, (b) suspense, and (c) action. The horizontal axis represents the shot number in the scene.

Figure 2 shows the histogram of the activity intensity values for three types of shots: (a) conversation, (b) action and (c) suspense.

3.2 Face Detection

Conversational scenes generally have shots with at least two humans. We utilize this cue and detect human faces in the video using the method proposed by Viola et al. [6]. We have found that [6] performs reasonably good for faces with different scales in the video. The shots containing faces are clustered together based on a 24-bin RGB histogram. Figure 3 shows human faces in some shots.

Fig. 3. Results of face detection in a scene.

3.3 Finite State Machines (FSM)

A Finite State Machine is defined as,

$$A = (Q, \Sigma, \sigma, q_0, F), \tag{4}$$

where Q is a set of states in the FSM and σ is the set of transitions. Σ contains the conditions for the transitions. q_0 is the initial state, and F is the set of accepting (final) state. In feature films, scenes are generally composed in accordance with the conventional film grammar. We have observed the following characteristics for three different categories of scenes:

- **(i) Conversational scenes:** low activity intensity, medium audio energy and multiple speakers.
- **(ii) Suspense scenes:** a long period of silence followed by a sudden eruption either in sound track or in activity intensity or both.
- **(iii) Action scenes:** intensive action activity for a certain number of shots.

We discuss three different FSMs which detect conversational, suspense and action scenes.

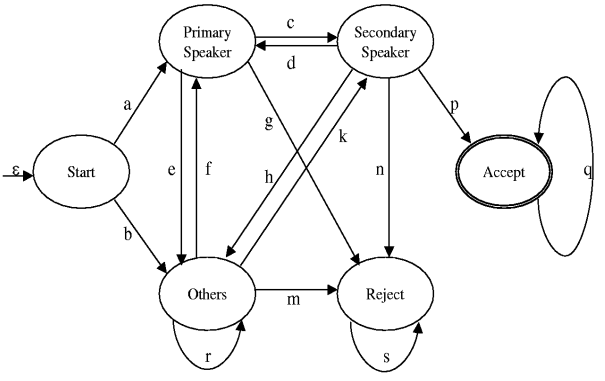


Fig. 4. Finite state machine for conversation scene detection. It consists of six states.

FSM for Conversation Scenes. Figure 4 shows a deterministic Finite State Machine for detecting conversation scenes. The FSM consists of six states: *Start*, *Primary Speaker*, *Secondary Speaker*, *Others*, *Reject* and *Accept*. Shots with high similarity that contain a face are clustered together. 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 *Accept* is reached, the scene is declared as "Conversation" scene. Otherwise it is declared as "Non-Conversation". In this FSM, $Q = \{Start, PrimarySpeaker, SecondarySpeaker, Others, Reject, Accept\}$, $q_0 = \{Start\}$ is the initial state and $F = \{Accept\}$ is the final state. The set of the transitions σ includes $\{\varepsilon, a, b, c, d, e, f, g, h, k, m, n, p, q, r, s\}$. The transition matrix for σ is shown in Table 1. The transition conditions Σ are:

- **a:** The first shot in the scene is a facial shot with low activity intensity. Results in the transition to the state *Primary Speaker*.
- **b:** The first shot in the scene is a non-facial shot with low activity intensity. Results in the transition to the state *Others*.
- **d** and **f:** The new shot is a facial shot with low activity intensity, and it belongs to the largest cluster. Results in the transition to the state *Primary Speaker*.
- **c** and **k:** The new shot is a facial shot with low activity intensity, and it belongs to the second largest cluster. Results in the transition to the state *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 neither to the largest cluster nor the second largest cluster. Results in the transition to the state *Others*.
- **g**, **n** and **m:** The new shot has high activity intensity. Results in the transition to the state *Reject*.
- **p:** The new shot is a facial shot with low activity intensity. It completes the accepting requirement of the FSM. Results in the transition 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*.

Table 1. Transition matrix for conversation detection. Columns represent "From" states, rows represent "To" states and "-" indicates no transition from one state to another.

σ	Start	Primary	Secondary	Others	Reject	Accept
Start	-	a	-	b	-	-
Primary	-	-	c	e	g	-
Secondary	-	d	-	h	n	p
Others	-	f	k	r	m	-
Reject	-	-	-	-	s	-
Accept	-	-	-	-	-	q

FSM for Suspense Scenes. We have observed that suspense scenes often have the following pattern. In the beginning, the scene is relatively silent and is followed by a sudden increase in sound energy. In many cases, it is also accompanied by abrupt camera and actor movements. Based on these observations, the FSM for detecting the suspense scenes have the following four states: *Start*, *Wait*, *Reject* and *Accept*. The state *Wait* represents the pre-action moments. After a period of *waiting*, the state is transferred to *Accept* if a sudden action shot is seen. The FSM rejects the scenes in which the sudden action happens before the predefined interval.

Similarly, the definition of the FSM for the classification of suspense scenes can be written in the general formula for the Finite State Machines. In this case, $Q = \{Start, Wait, Reject, Accept\}$ are the states. The initial state is $q_0 = \{Start\}$, and the final state is $F = \{Accept\}$. The transition set σ includes $\{\varepsilon, a, b, c, d, e, f, h\}$. The FSM is shown in Figure 5, and the corresponding transition matrix is shown in Table 2. The transition conditions are defined as follows:

- **a:** The first shot in the scene is a shot with low activity intensity. Results in the transition to the state *Wait*.
- **b:** The first shot in the scene is a shot with high activity intensity. Results in the transition to the state *Reject* if the action happens before a predefined time interval.
- **c:** The new shot is a shot with high activity intensity, and the waiting time is less than the required period. Results in a transition to the state *Reject*.
- **d:** The new shot is a shot with low activity intensity, and the waiting time is less than the required period. Loops at the state *Wait*.
- **e:** For any new shot, loops at the state *Reject*.
- **g:** The new shot is a shot with high activity intensity, and the waiting time is more than the required period. Results in the transition to the state *Accept*.
- **h:** For any new shot, loops at the state *Accept*.

FSM for Action Scenes. Action scenes in movies generally have very high action intensity, such as scenes containing explosions, chasing and fighting. To classify a scene as an action scene, the scene must have three or more shots with action intensity higher than a threshold. The FSM for detecting action scenes is shown in Figure

For action FSM, the state set Q has $\{Start, FirstAction, SecondAction, Non-Action, Accept(ThirdAction)\}$, where the initial state q_0 is $\{Start\}$, and the final state F is $\{Accept(ThirdAction)\}$. The transition set σ includes

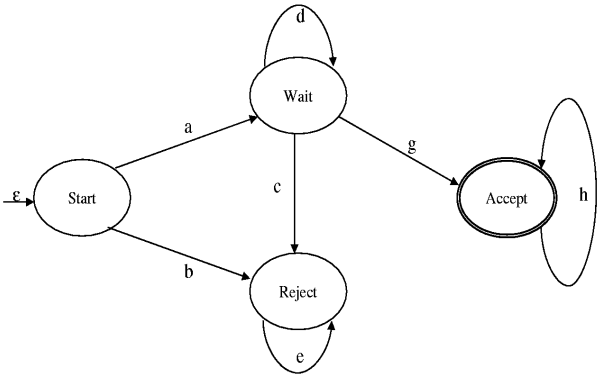


Fig. 5. Finite state machine for suspense scene detection. It consists of four states.

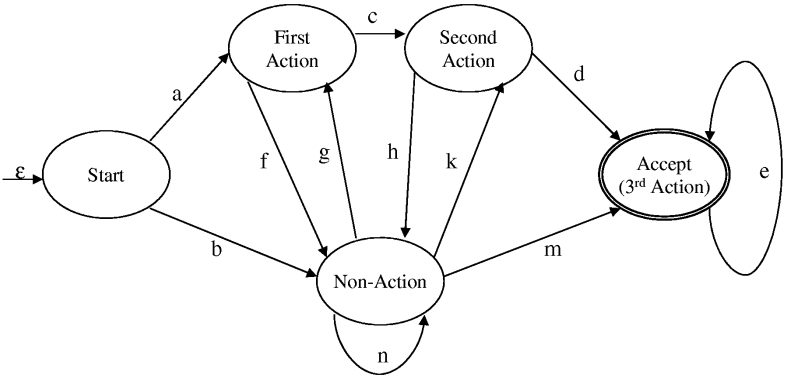


Fig. 6. Finite state machine for action scene detection. It consists of five states.

Table 2. Transition matrix for suspense scene detection. Column represent "From" states, rows represent "To" states and "-" indicates no transition from one state to another.

σ	Start	Wait	Reject	Accept
Start	-	a	b	-
Wait	-	d	c	g
Reject	-	-	e	-
Accept	-	-	-	h

$\{\varepsilon, a, b, c, d, e, f, g, h, k, m, n\}$. A *Previous State* attribute for a state q_i in the FSM and defined as the "from" state of the immediate transition before reaching state q_i. This is used for the determination of the outgoing transitions from state *Non-Action*. The transition matrix is shown in Table 3. The transition conditions are: 6.

– **a:** The first shot in the scene has high activity intensity. Results in the transition to the *First Action* state.

- **b:** The first shot in the scene has low activity intensity. Results in the transition to the state *Non-Action*. The *Previous State* is set to *Start*.
- **c:** The new shot has high activity intensity. Results in the transition to *Second Action*.
- **d:** The new shot has high activity intensity. Results in the transition to the state *Accept (Third Action)*.
- **e:** For any new shot, loops at *Accept (Third Action)*.
- **f:** The new shot has low activity intensity. Results in the transition to *Non-Action*. The *Previous State* is set to *First Action*.
- **g:** The new shot has high activity intensity. Results in the transition to the state *First Action*. The *Previous State* is set to *Start*.
- **h:** The new shot has low activity intensity. Results in the transition to the state *Non-Action*. The *Previous State* is set to *Second Action*.
- **k:** The new shot has high activity intensity. Results in the transition to the state *Second Action*. The *Previous State* is *First Action*.
- **m:** The new shot has high activity intensity. Results in the transition to the state *Accept (Third Action)*. The *Previous State* is *Second Action*.
- **n:** The new shot has low activity intensity. Loops at *Non-Action*.

Table 3. Transition matrix for action scene detection. Column represent "From" states, row represent "To" states and "-" indicates no transition from one state to another.

σ	Start	1st-Act	2nd-Act	Non-Act	Accept
Start	-	a	-	b	-
1st-Act	-	-	c	f	-
2nd-Act	-	-	-	h	d
Non-Act	-	g	k	n	m
Accept	-	-	-	-	e

4 Experimental Results

We have experimented with over 60 clips using the Finite State Machines for 3 categories of scenes. These clips are taken from 7 Hollywood movies including "The Others", "Jurasic Park III", "Terminator II", "Gone in 60 Seconds", "Mission Impossible II", "Dr. No", and "Scream". We also included a TV talk show, "Larry King Live" and a TV news program, "CNN Headlines". The feature movies cover a variety of genres such as horror, drama, and action. Each clip contains approximately 20-30 shots. Four human observers were asked to choose the most suitable label from three categories for each clip. Each clip was given a ground truth label with the category that the most human observers agreed upon. Thus, each clip is considered as a positive member of the category to which it is assigned. Observers were also asked to provide the most unlikely category for each clip. We used this information to label a clip as a non-member (or a negative member) for the unlikely categories.

To evaluate the performance of the proposed approach, two measures of accuracy were computed. These measures are precision and recall and defined as follows:

$$P_{pos} = \frac{M_{pos}}{D_{pos}}, \quad R_{pos} = \frac{M_{pos}}{G_{pos}} \tag{5}$$

and

$$P_{neg} = \frac{M_{neg}}{D_{neg}}, \quad R_{neg} = \frac{M_{neg}}{G_{neg}}, \tag{6}$$

where P_{pos}, R_{pos}, P_{neg} and R_{neg} are the precision and recall for positive and negative member detection. G_{pos} and G_{neg} are the ground truth. D_{pos} and D_{neg} are the detected positive and negative members. M_{pos} and M_{neg} are the numbers of the correctly matched positive and negative members.

There were 27 conversational scenes in the data set. The results achieved were 96.2% precision and 92.6% recall. For the other 25 non-conversational scenes, the precision was 92.0%, and the recall was 95.9%. The number of positive members of the suspense category in the data set was 12, with 15 non-member scenes. The precision and recall for the member detection was 100.0% and 93.8% respectively, and the precision and recall for the non-member clip was 91.7% and 100.0% respectively. In action scenes, we had 21 member clips and 29 non-member clips. The precision and recall for the positive members was 87.0% and 95.2% respectively. The precision and recall for the negative members are 96.3% and 89.7% respectively. The overall performance is summarized in Table 4. These results clearly demonstrate that a finite state machine can detect and classify video scenes into categories. Figure 7 shows some clips with the key frames of the shots in the scene.

(a) Conversation Scene: *007 - Dr. No*

(b) Suspense Scene: *Scream*

(c) Action Scene: *Terminator II*

Fig. 7. Three testing clips. The six representative key frames are displayed.

5 Conclusions

In this paper, we presented a novel framework for classifying video scenes into high-level semantic categories using deterministic Finite State Machine (FSM). The transitions in

Table 4. Precision and recall for conversation, suspense and action scene classification.

Scene Type	**Conversation**		**Suspense**		**Action**	
Accuracy	Positive	Negative	Positive	Negative	Positive	Negative
Precision	96.2%	92.0%	100.0%	93.8%	87.0%	95.2%
Recall	92.6%	95.9%	91.7%	100.0%	96.3%	89.7%

each FSM are based on the low and mid-level shot features. These features are robust and easily computable. We also incorporated face detection to cluster shots and used these clusters to determine the transitions of the FSMs. We demonstrated the usefulness of FSM for this task by experimenting on over 60 movie clips and achieved high recall and precision. In the future, we plan on exploring Finite State Machine to detect scene categories for entire movies.

References

1. B. Adams, C. Dorai, S. Venkatesh, *Novel Approach to Determining Tempo and Dramatic Story Sections in Motion Pictures*, ICIP, 2000.
2. Z. Rasheed, M. Shah, *Scene Detection In Hollywood Movies and TV Shows*, IEEE Computer Vision and Pattern Recognition Conference, Madison, Wisconsin, June 16-22 2003.
3. A. Yoshitaka, T. Ishii, M. Hirakawa, and T. Ichikawa, *Content-Based Retrieval of Video Data by the Grammar of Film*, IEEE Symposium on Visual Languages, 1997.
4. R. Lienhart, S. Pfeiffer, and W. Effelsberg, *Scene Determination Based on Video and Audio Features*, Proc. IEEE Conf. on Multimedia Computing and Systems, Florence, Italy, 1999.
5. Y. Li, S. Narayanan, C.-C. Jay Kuo, *Movie Content Analysis Indexing, and Skimming*, Kluwer Academic Publishers, *Video Mining*, Chapter 5, 2003.
6. P. Viola and M. Jones, *Robust Real-Time Object Detection*, International Journal of Computer Vision, 2001.