International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 7, September - 2012

Rule Based Scenes Retrieval Using Turing Machine

K. Karunakar
Associate Professor
CSE, SITE, India

B. Anand Kumar
Assistant Professor
CSE, Dilla University, Ethiopia

B. Yugandhar
Assistant Professor
CSE, SITE, India.

MRao. Batchanaboyina
Associate Professor
CSE, PACE, India.

Abstract

Instead of clustering video shots into scenes using low level image features, in this paper, we propose an improved rule-based model to extract simple dialog or action scenes those are formerly included in the movie. Through analyzing video editing rules and observing temporal appearance patterns of shots in dialog scenes of movies, we deduce a set of rules to recognize dialog or action scenes. Based on these rules, a Turing machine is designed to extract the present dialog or action scenes with formerly happened action or dialog scene in the movie. Our TM has been experimented on over 4 Movie clips and convincing results have been achieved.

Index Terms —Video Analysis and rules, Turing Machine, Action or Dialog Scenes, Data mining.

1. Introduction

Modeling of video requires identification and extraction of its components. Early video database systems segment video into shots, and extract key frames from each shot to represent it. Such systems have been criticized for two reasons: shots do not convey much semantics, and using key frames may ignore temporal characteristics of the video. Therefore, there have been several attempts [1, 2, 3] to cluster semantically related shots into scenes. However, current approaches only employ low-level image features, which may cause semantically unrelated shots to be clustered into one unit only because they may be "similar" in terms of their low-level image features. Furthermore, users may not be interested in the "general" scenes constructed in this way, but may focus on particular scenes. In particular, dialog and action scenes have special importance in video, since they constitute basic "sentences" of a movie that consists of three basic types of scenes [4]: dialogs without action, dialogs with action, and actions without dialog. Automatic extraction of dialog and action scenes from a video is an important topic for practical usage of video.

A given video clip may be (and commonly is) interpreted differently by different users. However, there is one viewpoint that is the most important: that of the video editor or director. From their viewpoint, a video is produced to express some concepts or stories that they want to communicate to the audience. The editing process follows certain rules that can be used in automatic extraction of scenes. Lei Chen Discussed in his paper, based on the video editing rules for dialog and action scenes, he propose a Finite State Machine (FSM) model to extract simple dialog or action scenes from movies. In this paper we propose the Turing Machine (TM) model to extract formerly Action or dialog scenes between same actors.

2. Related Work

The intention is to retrieve the action or dialog scenes that are happened in the past between the same set of actors. The intention behind this is audience should know why the couple of people fighting each other. So they need to know what the reason is for present action is or dialog scene. Present scene come back with flashback action or dialog scene. To show flashback scene the Turing machine pointer moves the previous scene based on particular transition.

Observation of a large number of video dialog and simple action clips reveals the existence of visual patterns, such as interleaving patterns of the appearance of the actors who are involved in a dialog. These visual patterns can form the basis for detecting dialog and action scenes. In order to extract these visual patterns, the video clips are analyzed from the point of view of how a dialog scene is produced. In this paper, we focus on the case where a dialog scene (DS) has at most two actors (a and b) in it. This assumption is made for two reasons. First, the rules for the positioning of actors and cameras are better understood and documented in movie literature. Second, the cases of two actors are easier to explain and demonstrate. The extension of our work to more actors is explained at the end of the paper. DS is composed of a set of shots.

A. Simple Dialog Scenes Patterns

In order to capture a dialog, two main factors must be considered:

- The spatial arrangement of the actors; and
- The placement of cameras to capture dialogs.

These two factors affect the appearance of actors in captured video shots. Through the analysis of actor arrangement and camera placement, we find that there are only three basic types of video shot patterns in a two person (call a and b) dialog scene:

- a shot in which only actor a's face is visible throughout the shot (Type A shot);
- a shot in which only actor b's face is visible throughout the shot (Type B shot); and
- a shot with both actors a and b, with both of their faces visible (Type C shot).

In addition to these, usually an insert or cut-away shot is introduced to depict something related to the dialog or not covered by those three types of shots. We use symbol # to represent it. These constitute *video type set* V= {A, B, C, #}.

B. Simple Action Scenes Patterns

The rules governing the actor arrangement and camera placement in simple action scenes (e.g., one-on-one fighting), are the same as those for producing simple dialog scenes. This is true even though, in an action scene, actors move rapidly and cameras follow the actors. Therefore, video shots in a simple action scene can also be classified into four types: A, B, C and # as defined above.

3. Proposed System

The basic idea is depicted in the block diagram, shows different steps first we give the video as an input then converting the film into symbols, then we propose the identification rules for action or dialog scenes based on the rules extract the scenes. The main intention of this paper is retrieving the scenes belongs to the same actors whose are participated in formerly action or dialog scenes.

A. Block Diagram of the proposed System

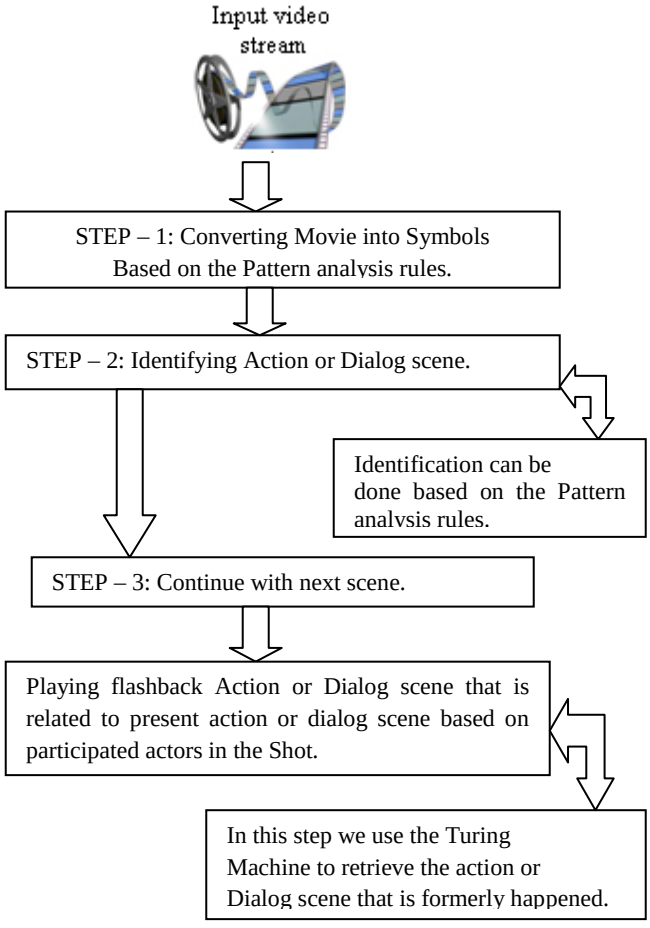


Fig. 1. Block Diagram for Proposed System

After a set of video shots are obtained from cameras that are used to film dialogs, the issue becomes how these shots can be used to construct a dialog scene to express a conversation. This is a challenging question for a video editor. However, there are some basic techniques that a video editor typically follows in constructing dialog

scenes [4, 5].

B. Editing Techniques to Construct a Dialog

Editing a dialog scene consists of two steps:

1. *Set Locale of the dialog scene:* In the first step, video editors set up the dialog scenario. The preference is for a scene that either consists of shots involving both actors (type C scene) or consists of shots that show alternating actors (i.e, either AB or BA), because these give the audience an early impression of who are involved in the dialog. During this setting up process, the basic building blocks of dialog scenes are constructed. We call these basic building blocks as Basic dialog scenes. An elementary dialog scene includes a set of video shots, and can itself be a dialog scene or be expanded to a longer dialog scene. The set of elementary dialog scenes are determined empirically, based on the analysis of editing rules used to establish dialog scenes and observations of dialog scenes of five movies1. As a result, we have identified eighteen types of elementary dialog scenes as depicted in Table 1 along with statistics about their occurrence frequency in the five movies under consideration.

Basic Dialog Scenes	Appearance Percentage
ABAB or BABA	38.15%
CAC or CBC	16.2%
ACC or BCC	3.3%
CAB or CBA	2.42%
ABC or BAC	17.21%
ABAC or BABC	2.42%
AA1 B1 C1BC	1.65%
BA1 B1 C1BC	1.55%
CA1 B1 C1BC	5.1%
AA1B B1C C1	5.2%
AB1B C1C A1	2%
BA1B B1C C1	3.2%
CA1B B1C C1	1.6%

Table 1. Statistical data on Basic dialog scenes

2. *Expanding the dialog scene with Flashback.* In the second step, each elementary dialog scene can be expanded by appending three types of shots. During this editing process, the basic rule that an editor uses is to give a contrast impression to the audience. For example, if the ending shot of one scene is an A type shot; usually a B type shot is appended to expand the scene. Similarly, the editor can append a C type shot as are establishing shot from time to time to remind the audience of the whole scenario surrounding the dialog scene.

Type of end shot in the scene	Types of shots that may follow
A	A1 or B or C
B	A or C
C	A or B or C

Table 2. Possible types of shots to be appended

C. Flashback Video Shot String

We introduce the concept of a Flashback Video shot string (FBVSS) to represent the temporal presentation sequence of different types of shots in a video. A FBVSS is a string which is composed of symbols from V Each symbol in V SS represents a shot in a video. The ordering of symbols in the string is from left to right or right to left, which

represents the shot presentation sequence.

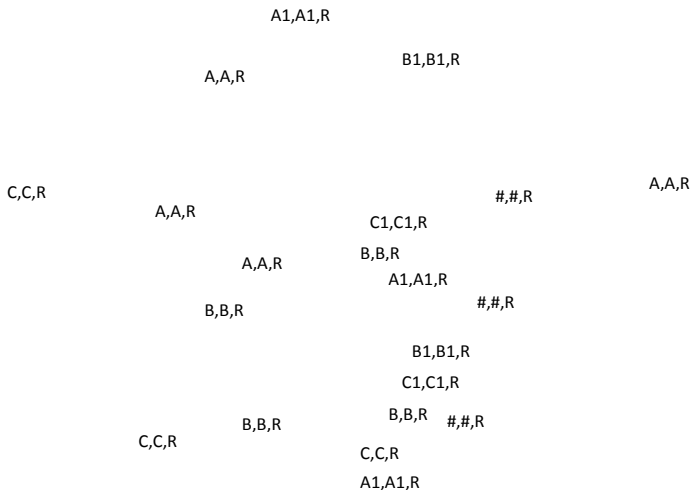
Based on the analysis of the above-discussed two editing steps, we Define a FBVSS of a dialog scene as a string whose prefix or postfix is one of the basic dialog scenes that can be expanded by the rules given in Table 2. The starting basic dialog scene classified a FBVSS as well. Consequently, there are thirteen types of FBVSSs corresponding to those types of dialog scenes. It is easy to prove that these are regular languages over set V. We do not give a complete proof due to lack of space, but the following is the proof of one of these cases, namely the FBVSS whose prefix is AA1BB1CC1. Proof of other cases are similar.{A},{A1},{B},{B1},{C} and {C1} are regular languages over V .{AA1BB1CC1} is a product of regular languages {A},{B} and {C}:{AA-1BB1CC1}={A} • {A1} • {B} • {B1} • {C} • {C1}, so {AA1BB1CC1} is a regular language over V, too. FBVSS that starts with AA1BB1CC1 includes string AA1BB1CC1 and all the strings which are expanded from AA-1BB1CC1 by appending A, A1, B, B1, C or C1 using the rules in Table2. Appending a shot to a scene is a concatenation operation (•). Therefore, by definition of a regular language [6], a FBVSS of a dialog scene that starts with ABAB is a regular language over V. By taking the union of the eighteen types of FBVSSs, we again obtain a regular language over set V. Therefore, V SSs that are used to represent the temporal appearance patterns of video shots in dialog scenes are regular languages over set V.

D. Turing Machine to Extract FBVSSs of Dialog Scenes

Since FBVSSs of dialog scenes are regular languages, the next issue will be how to automatically extract the FBVSSs which correspond to dialog scenes from FBVSSs of the whole video. In there words, how to extract specified regular languages from FBVSSs? In this paper, we propose a Turing machine (TM) model to extract dialog scenes from videos. Note that we are not using the TM to determine whether a language is a regular language over V, but constructively to extract those parts of a FBVSS that form regular languages with certain properties. In our proposed TM, a VSS is used as an input to the TM. A state is used to represent the status after a number of shots have been processed. An edge between states will determine an allowable transition from current state to another state under a labeled condition. The label of the arc is a symbol which is used to represent a type of shot. A sub-string of the FBVSS will be extracted by the TM if and only if there exists a path from initial state to one of final states. The symbols on the path correspond to sequence of the shots in that sub-string of FBVSS. Figure 2 shows the transition diagram of our proposed TM which is used to extract FBVSSs of simple dialog scenes between two actors.

E. Extract Action Scenes

Since a video editor follows similar rules that are used to construct dialog scenes to compose simple action scenes, temporal appearance patterns of video shots in simple action scenes are similar to those of dialog scenes. The TM model discussed above is also suitable for extracting simple action scenes (one-on-one fighting). However, in order to give the audience a different feeling between action scenes and dialog scenes, several other techniques are used to enhance visual effects. These techniques involve manipulating the length of a shot and combining static and moving cameras, etc. Among these, the length of a shot is an important factor to express visual effects of an action scene. In our approach, the average length of shots in a scene will be used to differentiate between a dialog scene and an action scene.



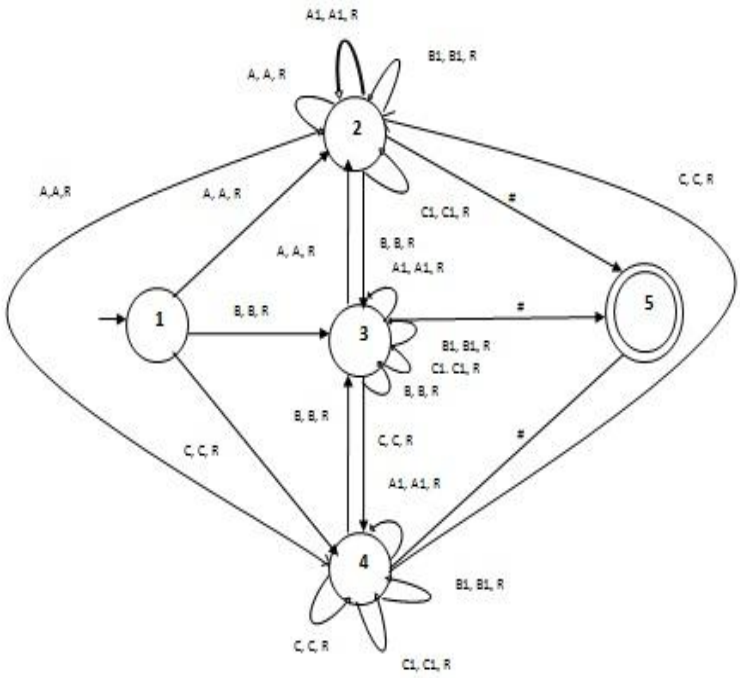


Fig. 2. A TM extracts *FBVSS*s of dialog scenes between actor a and actor b

δ	A	B	C	A_1	B_1	C_1	#	B
→ **1**	(A,R,2)	(B,R,3)	(C,R,4)	-	-	-	-	-
2	(A,R,2)	(B,R,3)	(C,R,4)	(A_1,R,2)	(B_1,R,2)	(C_1,R,2)	(#,#,R)	-
3	(A,R,2)	(B,R,3)	(C,R,4)	(A_1,R,3)	(B_1,R,3)	(C_1,R,3)	(#,#,R)	-
4	(A,R,2)	(B,R,3)	(C,R,4)	(A_1,R,4)	(B_1,R,4)	(C_1,R,4)	(#,#,R)	-
***5**	-	-	-	-	-	-	-	-

Table 3. Transition Table for TM Model

4. Experiment Results

In this section, we present the results of some extraction experiments that were conducted using our TM model. Four different movies are used in our experiment (Table 4), which is first segmented into shots and appearances of actors are manually marked. Our focus is on retrieval precision and recall, which are defined identical to their use in the information retrieval literature. Precision measures the proportion of correctly recognized scenes, while recall measures the proportion of scenes that are recognized. Tables 5 and 6 show the result of extracting simple dialog scenes with its flashback and simple action (one-on-one fighting) scenes with its flashback scene between same set of actors respectively. Table 5 shows that our model can achieve high precision and recall in extracting dialog scenes with flashback from movies. There is an interesting fact in Table 5 that the TM model achieves better results on the movie "3 Idiots" compared to the other three movies. As shown in Table 6, and as our analysis showed, "3 Idiots" is completely composed of dialog scenes without any action scenes. This ensures that we do not falsely detect a dialog scene as an action scene or vice versa. In the movie "Wrong Turn", the precision is low, because several action scenes are mixed with dialog scenes. This is an artifact of the fact that we use the same Turing Machine (which implements the same set of rules) to detect both types of scenes extending with flash backs, and in some action scenes in this movie, long shots are used to show the action effects, which leads our model that uses the average shot length to differentiate dialog and action scenes to misclassify these action scenes as dialog scenes.

After we extract simple dialogs from movies, we can easily retrieve dialog scenes involving three or more actors. This can be achieved by finding pairs of dialog scenes with a common actor and overlapping durations. Table 7 shows the performance of this approach in detecting multi-actor dialog scenes in the four movies under consideration.

Movie Title	Genre	Year	Duration (Min)	No. Shots
I Am Legend	Action	2001	90	1256
Wrong Turn	Action	2007	96	1350
3 Idiots	Comedy	2009	120	1131
Slum Dog Milliner	Comedy	2009	120	1050

Table 4. The experiment data

Movie Title	No. Detected Dialogs	Precision (%)	Recall (%)
I AM LEGEND	95	89.47	96.60
Wrong Turn	154	80.52	90.51
3 Idiots	205	94.17	98.76
Slum Dog Milliner	195	91.79	97.28

Table 6. Dialog scenes extracted by the TM

Movie Title	No. Detected Actions	Precision (%)	Recall (%)
I AM LEGEND	35	84	84
Wrong Turn	54	81	89
Slum Dog Milliner	22	91	84

Table 7. Action scenes extracted by the TM

5. Conclusion and Future Work

In this paper, based on the analysis of video editing techniques, a set of rules on the temporal appearance patterns of shots are deduced. A TM is designed based on these rules to extract simple dialog or action scenes with flashback. The experimental results show that our model can efficiently extract dialog or action scenes with formerly action or dialog scenes from movies and with simple dialog scenes extracted from movies. Our TM model is a rule based model, it will be very suitable for online query processing. As we know, audio is an important feature for video analysis. Our future work will extracting video clips in films based on audio. Our model achieve higher accuracy and to extract more types of semantic scenes.

There is limited related work in the area. Yoshitiaka et al. [7] propose an algorithm to extract scene semantics (conversation, tension rising, and action) based on a grammar of the film. However, in their approach, only the repetition of similar shots (A − B − A0− B0) is employed to detect conversion scenes. Also Lienhart et al. [8] develop a technique to extract dialog scenes with the aid of an face detection algorithm. Lei Chen and M. Tamer O¨ zsu [9] propose a FSM to extract action or dialog scenes based on grammar of the films. However, they only extract dialog or action scenes which show shot/reverse shot patterns. Compared to all of these approaches, our method has following advantages:

- We can, in addition to shot/reverse shot dialogs, detect single shot dialogs, dialogs with insertions and cuts, and dialogs with shot/reverse shots and recovering shots.
- In addition to action or dialog scene, extracting the flashback scene between same actors by this audience know the reason behind this action shot.
- Our model is rule based, which is very suitable for on-line content based query processing.

6. References

[1] M. M. Yeung and B. L. Yeo, "Time-constrained clustering for segmentation of video into story units," in *Proceedings of 13th International Conference on Pattern Recognition*, 1996, pp. 375–380.
[2] Y. Rui, T. S. Huang, and S. Mehrotra, "Exploring video structure beyond the shots," in *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, 1992, pp. 237–240.
[3] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automatically segmenting movies into logical story units," in *Proceedings of International Conference on Visual Information Systems*, 1999, pp. 229–236.
[4] D. Arijon, *Grammar of the Film Language*, Focal Press, 1976.
[5] S. D. Katz, *Film Directing Shot by Shot Visualizing From Concept to Screen*, Michael Wiese Productions1991.
[6] J. L. Hein, *Theory of Computation: An introduction*, Jones and Bartlett Publishers, 1996.
[7] A. Yoshitaka, T. Ishii, and A. Hirakawa, "Content based retrieval of video data by the grammar of film," in *Proceedings of IEEE Symposium on Visual Languages*, 1997, pp. 310–317.
[8] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Scene determination based on video and audio features," in *Proceedings of International Conference on Visual Information Systems*, 1999, pp. 685–690.
[9] Lei Chen and M. Tamer O¨ zsu, "Rule-Based Scene Extraction From Video".
[10] Lei Cheny, Shariq J. Rizviz and M. Tamer Ozsuy, "Incorporating Audio Cues into Dialog and Action Scene Extraction".