ACTION SEQUENCE DETECTION IN MOTION PICTURES

BART LEHANE, NOEL E. O'CONNOR, and NOEL MURPHY
Centre for Digital Video Processing
Dublin City University
E-mail: {lehaneb, oconnorn, murphyn}@eeng.dcu.ie

This paper describes an approach for automatically detecting action sequences in movies. We examine the filmmaking conventions that are inherent in action sequences and use these as a basis for our analysis. A system of low- mid- and high-level features is presented that combines pure digital video analysis at the low- and mid-levels, with high-level filmmaking knowledge. A state machine is used to combine these features and detect the action sequences. The overall system is designed so that the analysis can be used to detect different types of scenes, but in this paper we focus on action sequence detection.

1 Introduction

In our previous work [1] we demonstrated how it is possible to use low-, mid- and high-level visual features to detect the presence of dialogue sequences in a movie. We closely examined the directorial and editing conventions followed by filmmakers when creating dialogue sequences and, based on this, targeted a small number of visual features (e.g. camera motion, visual similarity etc.) that we then combined using a state machine to detect dialogue sequences. In this paper we extend on the same ideals of basing our analysis on well-defined filmmaking conventions, but now focus our attention on action sequences.

There has been substantial work in detecting scene changes in digital video. Kender et al [2] and Yeo et al [3] use a memory-based approach to scene change detection that measures the visual distance between previous shots and the current shot. Huang et al [4] and Sundaram et al [5] used a combination of both video and audio to assist in the determination of scene breaks. This approach is based on the idea that the audio should change as well as the video in any scene changes.

The aim of this paper is to detect events rather than scenes. A single scene may contain a number of these events. Previous work in this area includes Nam et al [6] who created a *spatio-temporal activity measure* for each shot in order to find violent sequences in movies. Chen et al [7] use visual and audio cues to detect both dialogue and action scenes where the term action scene is used to address one-on-one fighting only. Methods such as [8] produce 'skims' that display a shortened version of the movie, with the aim being to show as much of the action as possible.

There are many potential applications of this type of event detection. A film summarisation system is envisaged, where users can quickly and easily

browse the content of a film. This could be used in an on-line rental context, where users can preview certain portions of a film before renting. Similarly, in a large movie database, the retrieval of particular sequences would be supported.

2 Film Syntax for Action Sequences

It is a general filmmaking concept that is necessary to keep the audience's attention at all times. This means that at different times in the film, directors and editors have to employ different styles adapted to the events taking place. For example, in a conversation sequence, it is necessary for the audience to comfortably view the characters [12]. In an action sequence however, the objective of the director is somewhat different, since his/her main aim is to excite the viewer, and to make sure that the audience cannot relax, as they should be engrossed in the action taking place. Excitation in action sequences is typically accomplished by movement within shots, movement of shots, and variation in the length of shots. Pans, tilts, and zooms are used to follow characters moving within shots. The camera itself moves to record these shots. Using pace and making shots shorter also helps to increase the excitement of the sequence [11].

3 Action Sequence Detection System

3.1 System Overview

The proposed action sequence detection system uses a combination or low-, mid-, and high level features as shown in Figure 1. The low-level features produce information about the video that is not necessarily of any interest to the user, but that is useful for higher level processing. Mid-level blocks produce information that is at a shot level. The high-level block accepts information from the low-and mid-level blocks and uses this to make a decision about the content under consideration during a given time interval. It produces the output retrieval unit, in this case the action sequence.

3.2 Shot Boundary Detection

Determining the shot boundaries is a key essential step prior to performing shot-level feature extraction and any subsequent scene-level analysis. To this end, a histogram-based shot boundary detection approach is used in order to detect boundaries and extract keyframes [9][1].

3.3 Motion Activity Analysis

This block uses the P-Frame motion vectors of an MPEG-1 video sequence in order to compute a global motion intensity measure for each shot based on calculating the standard deviation of the motion vectors in each P-Frame [13].

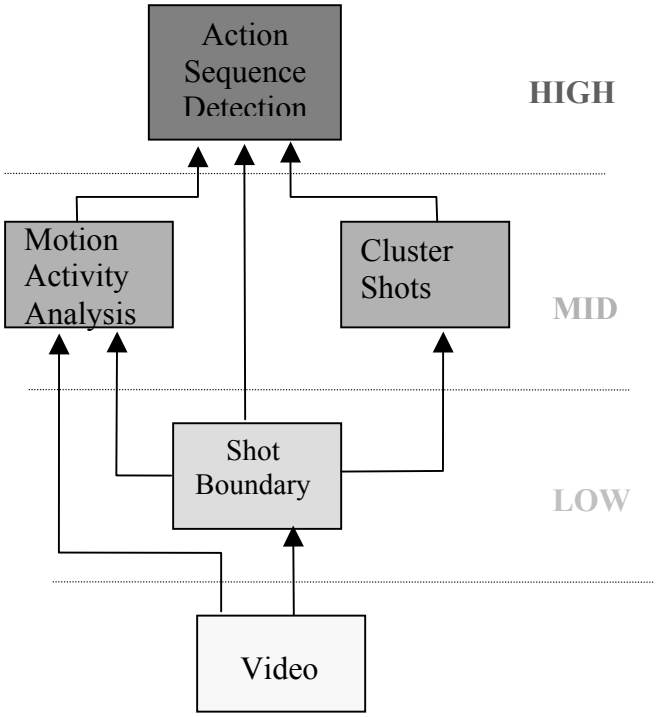


Figure 1: Action Sequence Detection System

3.4 Shot Clustering

The Shot Clustering block clusters visually similar shots that are temporally close together. This is based on the method used by Yeung et al [10]. The distance between the clusters is calculated based on the average color histogram of keyframes in the cluster.

Shot clustering algorithm:

1) Make N clusters, one for each shot.
2) Stop when the difference between 2 clusters is greater than a predefined threshold.
3) Find the most similar pair of clusters, R and S within a specified time constraint.
4) Merge R and S (More specifically merge S into R).
5) Go to step 2.

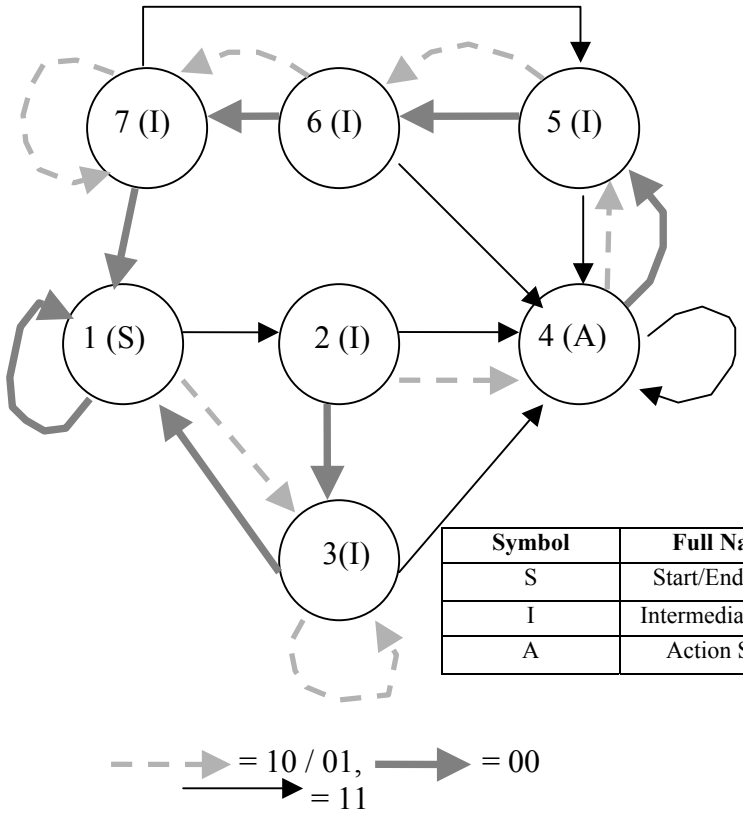
The time-constraint ensures that only shots that are less than 2000 frames (just over a minute) apart can be merged.

3.5 Action Sequence Detection

This high-level block accepts inputs from both the motion activity and clustering mid-level blocks, as well as the shot boundary low-level block. Its aim is to detect sequences of shots that have the characteristics of an action sequence for

a given set of inputs. It uses a two-pass method to do this. Firstly, a state machine was created to look for sequences that match the structure of action sequences as laid down by directorial and editing conventions. Then these Potential Action Sequences (PAS) are either accepted or rejected as action sequences based on the clustering input.

The state machine is illustrated in Figure 2. An essential part of any action scene is an escalated amount of motion and a quickening shot cut rate in order to grab the attention of the viewers. The state machine is designed to look for sequences of shots in which temporally short shots with high motion activity values are dominant. In figure 2, the thin black arrow indicates the action that the state machine takes when the shot under examination has both high motion and a short length (i.e. '11'). The red arrow shows the action that the state machine takes if the shot has both a long temporal length and low motion ('00'). The dashed blue arrow shows what happens when the shot has either a short length ***or*** high motion, but not both ('10' / '01').



Symbol	Full Name
S	Start/End point
I	Intermediate state
A	Action State

Figure 2: State machine for detecting PAS`s

The functionality of the state machine is as follows. It begins in state 1 (start point). If it encounters shots that are consistent with those of an action

sequence, then it tends towards state 4 (action) where it stays until it encounters shots that are dissimilar to action sequence shots and then reverts to state 1 (end point). Of course, in any action sequence, there will be shots that do not exactly follow the high motion and short length pattern, so it is necessary to have states that allow for these shots without termination. Similarly, there will be shots in other parts of the movie that happen to have a short length and/or high motion activity, but are not part of an action sequence. The intermediate (I) states are responsible for ensuring that the state machine doesn't classify these false positives as action sequences.

Once a potential action sequence is generated, a post-processing step is carried out to detect false positives. In an action sequence there should be very little shot repetition as the camera should change its viewpoint. In a dialogue sequence there is significant shot repetition, as the director will show shots of the same characters from the same angle to let viewers relax and concentrate on the dialogue. In an action sequence, however, the objective of the filmmakers is for the viewers to be excited, shocked, tense etc. As such, directors ensure that each new shot gives the viewer something different to understand. In addition, due to the nature of many action events, the focus of interest (a car in a car chase for example) will shift rapidly to different locations, which will again lead to low shot repetition. For this reason, the cluster to shot (C:S) ratio of each PAS is examined. If there is little shot repetition, this ratio will be quite high. We use an empirically chosen threshold to decide if the PAS is accepted as a true action sequence.

4 Results / Experiments

Generating an experimental ground truth for action sequences is problematic. The term 'Action Sequence' can cover a large array of events, from a fight, to a car chase, to a sports event, to a battle etc. According to filmmaking convention, action sequences are characterised by their use of pace, movement and subjective camera placement and movement. Intensification is achieved by varying the length of the shots. Conventionally, it means shortening the shots as the sequence approaches the climax. [11]. Thus, in order to make a ground truth of action sequences a certain amount of user discretion is required. The action sequences in a number of movies were manually marked base on the following rough guidelines:

- There must be some type of action event.
- There should be very little (if any) conversation between characters during an action scene

The results of the action sequence detector are presented in Table 1. Recall values tend to be high, in some cases returning all of the marked up action sequences. Precision values, on the other hand, are lower, usually due to short sequences that have the same characteristics as action sequences thereby giving false positives. Also, montage sequences are often shot in the same way as action sequences which, again, leads to false positives.

Table 1: Action sequence detection results.

Film Name	Num. Action Sequences	Num. Detected	Precision	Recall
Snatch	21	21	70 %	100 %
Kill Bill	17	15	65 %	88 %
Reservoir Dogs	5	4	40 %	80 %
Dumb and Dumber	14	14	60 %	100 %
Total	57	54	62 %	95 %

The type of action sequence was also recorded while creating the ground truth, and has been divided into a number of categories as shown in Table 2. Of these, "fights/arguments" are the most common type of actions sequence. Although a definition of an action sequence could be drawn from this list (as in 'Any sequence that contains a fight, a shooting, a chase a robbery or an accident') it is not a complete list and many more films will need to be viewed before an accurate definition could be derived in this manner. For example, there are no sports sequences in any of these films (apart from a boxing match, which comes under the 'Fight/Argument' category), and these sequences should surely also be considered.

5 Conclusions

In this paper we presented an approach to detecting action sequences in movies based on filmmaking convention. One of the main problems we encountered was how to test the system. Due to the ambiguous nature of the term 'Action Sequence' we had to apply our own understanding to generate a ground truth. Our ground truth would undoubtedly be different were it marked up by different individuals, as a strict definition of an action sequence could not be made. A more consensual ground truth will be generated in the future based on merging the mark-up of multiple users provided with more formal guidelines. The audio track was not considered in this preliminary work, and characteristics such as speech/music discrimination would be quite useful to the analysis. Also, we aim to apply our techniques of action sequence detection to media other than movies. Fictional television programmes or documentaries should also be considered.

Table 3: Categories of action sequences manually marked up.

Film Name	Num. Action Sequences	Fight/ Argument	Shooting	Chase/ Race/ Getaway	Robbery	Accident/ Crash	Other
Snatch	21	8	3	3	3	2	2
Kill Bill	17	14	1	2	0	0	0
Reservoir Dogs	5	2	0	3	0	0	0
Dumb and Dumber	14	4	2	4	0	4	0

6 Acknowledgements

The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged. This material is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA.

7 References

[1] B. Lehane, N. O'Connor and N. Murphy. Dialogue scene detection in movies using low and mid-level visual features. To appear in proceedings of *International Workshop on Image, Video, and Audio Retrieval and Mining*. Univ. Sherbrooke, Oct 25-26 2004.

[2] John R. Kender and Boon-Lock Yeo. Video Scene Segmentation Via Continuous Video Coherence. Proc. *CVPR '98*, pp 367-373, June 1998.

[3] B.-L. Yeo and B. Liu. Rapid scene analysis on compressed videos. *IEEE Trans. Circuits Syst. Video Technol. 5*, 6 (Dec. 1995), 533–544.

[4] Jincheng Huang, Zhu Liu, and Yeo Wang. Integration of audio and visual information for content-based video segmentation. *IEEE Int'l Conf. Image Processing (ICIP98), Special Session on "Content-Based Video Search and Retrieval"*. Oct. 1998. Chicago.

[5] Hari Sundaram and Shih-Fu Chang. Determining Computable Scenes in Films and their Structures using Audio-Visual Memory Models. *ACM Multimedia 2000*, Oct 30 - Nov 3, Los Angeles, CA.

[6] Jeho Nam, Masoud Alghoniemy and Ahmed H. Tewfik. Audio-Visual content based violent scene characterization. The international conference on image processing (ICIP). October 04 - 07 1998. Chicago.

[7] L Chen, S. J. Rizvi, and M. T. Özsu, "Incorporating Audio Cues into Dialog and Action Scene Extraction", In *Proceedings of SPIE Storage and Retrieval for Media Databases*, San Jose, CA, January 2003, pages 252-264.

[8] Hari Sundaram and Shih-Fu Chan. Condensing computable scenes using visual complexity and film syntax analysis. *IEEE Conference on Multimedia and Exhibition*, Tokyo, Japan, Aug. 22-25, 2001

[9] Paul Browne, Alan F. Smeaton, N. Murphy, N. O`Connor, S. Marlow, and C. Berrut. Evaluation and Combining Digital Video Shot Boundary Detection Algorithms. *IMVIP 2000 - Irish Machine Vision and Image Processing Conference*, Belfast, Northern Ireland, 31 August - 2 September 2000.

[10] M. Yeug and B.-L Yeo, Video visualization for compact presentation and fast browsing of pictorial content, *IEEE Trans. Circuits Syst. Video Technol. 7*, 5 (Oct. 1997), 771–785

[11] Ken Dancyger, The Technique of Film and Video Editing: Theory and Practice, Focal Process, 2001

[12] David Bordwell and Kristin Thompson, Film Art: An Introduction, McGraw-Hill, 1993.

[13] B.S. Manjunath, Philippe Salember and Thomas Sikora. Introduction to MPEG-7, Multimedia content description language. John Wiley and Sons ltd. 2002.