An Innovative Model of Tempo and Its Application in Action Scene Detection for Movie Analysis

Anan Liu[1,2], Jintao Li[2], Yongdong Zhang[2], Sheng Tang[2], Yan Song[2], Zhaoxuan Yang[1]
[1]School of Electronic Engineering, Tianjin University, Tianjin, China
[2]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
liuanan@ict.ac.cn

Abstract

In this paper, we present an innovative model of tempo and its application in action scene detection for movie analysis. For the first time, we clearly propose that tempo indicates the rhythm of both movie scenarios and human perception. By thoroughly analyzing both aspects, we classify the factors of tempo into two sorts. The first is based on the film grammar and we use the low level features of Shot Length and Camera Motion to describe filmmaking by directors. The second is based on the human perception and we originally propose the information measure for perception depending on the cognitive informatics, a newly emerging and significative subject. With the information in both visual and auditory modalities, the low level features of Motion Intensity, Motion Complexity, Audio Energy and Audio Pace are integrated for the formulation of information to describe the viewers' emotional changes to continuously developing storyline. With both aspects, tempo is defined and tempo flow plot is derived as the clue of storyline. On the basis of video structuralization and movie tempo analysis, we build a system for hierarchical browse and edit with action scene annotation. The large-scale experiments demonstrate the effectiveness and generality of tempo for action movie analysis.

1. Introduction

The movie industry is an active producer of video. Every year about 4,500 movies are released around the world spanning over approximately 9,000 hours of video [1]. Therefore, with such a massive amount of information, there is a great need of the way in which we can browse the movies conveniently and retrieve video clips with special semantic concepts.

Early researches in content-based video analysis mainly focus on video structuralization and retrieval and the video data are news video, sports video and so on because they generally have some structure characteristics and the domain-based knowledge facilitates the video analysis. However, the retrieval results using low-level visual features are far from satisfaction. This fact has motivated the effort to perform semantic analysis. With the great success of semantic analysis in the video mentioned above, the research objects have extended to movies. Due to the complex storyline and potential film grammar, it is very difficult for researchers to analyze movies as they have done in other kinds of video. The earliest work specialized in semantic analysis for movies was presented in [2] by Nuno Vasconcelos et al. In this paper, the average video shot activity and the duration are used as features for the categorization of movies according to the violence. Jeho Nam et- al in [3] exploit multiple audio-visual signatures to create a perceptual relation for conceptually meaningful violent scene identification. Although these researches have obtained some success, they only put the former semantic analysis methods on movies and do not mine their characteristics. The most representative work for movie content analysis is presented in [4-7] by Brett Adams et al. In these papers, for the first time, they bring in the film grammar and present an original computational approach for extraction of movie tempo for deriving story sections and events that convey high level semantics of stories portrayed in video. Due to the limitation of using only a single modality, Lei chen et al in [8] incorporate audio and visual cues for dialog and action scene extraction. They also implement the method for making action movie trailers in [9]. Moreover, Hsuan-Wei Chen et al in [10] use the features of audio energy to modify tempo proposed by Brett Adams et al. With the improved tempo, they present a method for movie segmentation and summarization.

Although some research has been done on tempo model and movie analysis, there exist two important problems. On one hand, some new low level features are used to improve the tempo model without scientific basis. For example, Chen et al in [10] present the simple linear combination of shot length, motion intensity and audio importance score for tempo. However, there is less analysis for the relationship and additivity of these factors. On the other hand, researchers paid much attention on the elements of film grammar and neglected the influence of human perception for tempo computation because for effectively analyzing movies, the tempo model should include the

changes of both movie scenarios and human perception considering the interaction of viewers and movies themselves. Therefore, depending on the researchers' previous work and the instruction of cognitive informatics, a newly emerging subject, we present the innovative computational methodology for the model of tempo. The main contributions of our work are threefold: the integration of the elements of film grammar and human perception, the proper measure and combination of these elements and a system for hierarchical browse with action concept annotation on the basis of video structuralization and movie tempo analysis.

The remainder of the paper is organized as follows. In Section 2, we briefly introduce the influence of film grammar and human perception for movie analysis. Then, we specifically illustrate the computational methodology of the innovative model of tempo in Section 3. In Section 4, a system for hierarchical browse and edit with action concept annotation is introduced. The experimental results are presented in Section 5. At last, the conclusions and future work are stated in Section 6.

2. Film grammar and human perception

For movie content analysis, film grammar, the force that goes toward shaping movies, should be deeply analyzed. Film grammar is defined in [14] as being "comprised of a body of 'rules' and conventions" that "are a product of experimentation, an accumulation of solutions found by everyday practice of the craft," and results from the fact that films are crafted, built, shaped to convey a purpose. Depending on cinematic convention, there are two essential factors for movie content analysis. One factor is related to the editing technique of montage. Montage is an idea of film editing, deriving from the concept that there should be contrast between two different shots that are independent of each other [15]. With this method, directors bring special meaning to viewers with conspicuous contrast and change. The other factor is related to camera techniques. The elements, such as distance between cameras and characters, the shooting angle and the motion of cameras, can also delivery special meaning. With these factors of film grammar, directors aim to express potential semantic meanings to the viewers.

As a neurobiological concept, perception means the awareness of the elements of environment through physical sensation in [11]. Human beings receive and process certain types of external stimuli, including ocular and aural stimuli, and change the concentration of mental powers on the storyline. As for human perception, people usually focus on the high intensity and conspicuous changes both in audio and visual modalities. For example, an explosion or dramatic object motion can attract viewers' much attention. Therefore, for effectively analyzing the movie content, we also need formulating human emotional descriptors to quantitatively represent the change of human perception from the viewers' standpoint.

3. The model of tempo

Tempo carries with it the important notions of time and speed and its definition reflects the complexity of the domain to which it is applied [7].In this section, we will present the formulation of the tempo model from viewpoints of both film grammar and human perception. Because the tempo model with film grammar descriptor has been detailed in [4-7], we only simply introduce it. Then we will focus on the representation for human perception descriptor. With these efforts, we formulate a novel tempo function to depict the potential clue of the storyline.

3.1. Film grammar descriptor

Depending on the analysis in Section 2, film grammar descriptor consists of two elements. With the editing technique of montage, the director controls the speed at which a viewer's attention is directed and thus impacts on her appreciation of the tempo of a piece of video [4]. This feature can be represented by Shot Length. The continuous shots with short length in temporal domain indicate that there are more changes of scenarios and therefore a high tempo is delivered. Besides, with camera technique, especially camera motion, pan, tilt, zoom and so on, the director influence viewers' emotional feelings and attract their attention. Therefore, the magnitude of camera motion is computed to represents this element.

As for a video, we implement the shot boundary detection and camera motion estimation as presented in [20]. Then film grammar descriptor, FG(n), is formulated as follows:

$$FG(n) = \frac{\alpha(med_s - s(n))}{\sigma_s} + \frac{\beta(m(n) - \mu_m)}{\sigma_m} \qquad (1)$$

where s denotes shot length, m denotes motion magnitude, n is shot number, *meds* means the shot median, and μ and σ respectively denotes the mean and standard deviation of these features. Here, α and β are set with 0.5 assuming that both elements contribute equally to film grammar descriptor.

3.2. Human perception descriptor

Quantitative representation of human perception with scientific basis has been a problem unsolved ideally for a long time. However, cognitive informatics, the newly emerging subject, has founded the relative theoretical foundation and proposed an advisable method. Cognitive informatics is a disciplinary that studies the internal information processing mechanisms of the brain and their engineering applications via interdisciplinary approach [12]. In [12, 13], Wang proposes that information is a more proper measure for human perception. With calculating the information of both visual and auditory modalities, we can

represent the human perception descriptor quantitatively.

3.2.1 Visual information

As for movies, drastic and complex motion of objects will have great impact on viewers' emotional feelings and attract them greatly. Therefore, as visual features, motion intensity and complexity can be used for calculation of visual information.

On the step of video preprocessing, motion vectors value of 16*16 macroblocks are extracted from the MPEG-1 compressed video. Then we calculate motion based visual features as follows:

(a) Motion Intensity:

We use motion activity descriptor defined in MPEG-7 to represent motion energy. As for each P frame (for less computational complexity, we only process P frame for each video), the motion intensity (MI) of the $(i,j)^{th}$ macroblock, $MI_{MV}(i,j)$, is defined as:

$$MI_{MV}(i,j) = \sqrt{x_{i,j}^2 + y_{i,j}^2} \quad (2)$$

where $(x_{i,j}, y_{i,j})$ is the motion vector of the $(i,j)^{th}$ macroblock.

The motion intensity of each P frame is calculated by accumulating the motion intensity of macroblocks in one frame. Then we average MI of all the P frames in one shot for the calculation of shot-level motion intensity.

(b) Motion Complexity:

It is perceptive that more complex the motion is, more attention will be paid by viewers. Motivated by cognitive informatics, we use entropy of motion orientation to model this attribute.

Firstly, we calculate the orientation histogram with N bins to represent the phase distribution of motion vectors. Then, the entropy of motion orientation of the W^{th} P frame is calculated as follows:

$$MC(P_W) = -\sum_{n=1}^{N} h(n) Log(h(n)) \quad (3)$$

where MC denotes motion complexity and h(n) means the n^{th} bin of orientation histogram. Then shot-level Motion Complexity of each shot, MC^{AVE}, is calculated by averaging the MC of all the P frames in one shot.

As for viewers, highly dispersed motion vectors mean more kinds of motion existing in one frame. Thus, viewers will be likely to pay more attention on it. Comparatively, although the motion with single orientation and high intensity can attract viewers, people only need paying less attention on it due to the less complexity of visual content. In conclusion, the maximum motion complexity will be got when motion vectors disperse equally. This assumption just accords with the principle of maximum possible entropy in [22]. Therefore, the model in equation (3) is proper for the calculation of motion complexity.

From the illustration of motion intensity and complexity above, we can see that motion intensity represents energy and motion complexity reflects information. Therefore they belong to two categories and are not additive. However, the high values of them are all positive to represent the large visual information and indicate the high tempo. Therefore, we integrate them as follows to formulate the visual information, VI, by regarding motion intensity as the weight of motion complexity:

$$VI = MI * MC \quad (4)$$

3.2.2 Auditory information

As for movies, the accompanying sound is also very important to express the semantic meanings. Especially, large volume and high pace of audio influence viewers with strength and haste. Therefore, audio energy and audio pace are calculated for the constitution of auditory information.

(a) Audio Energy

Referring to [17], we calculate the frame-based short time energy to represent Audio Energy (AE). Then the short time energy is averaged for each shot to represent the importance of accompanying sound.

(b) Audio Pace

After AE is calculated, we use N_{AP} to represent the number of audio peaks that are greater than a threshold Th_{AP}, an empirical value. N_{AP} is further normalized by the total number of audio frame in one shot, N_{Total}. Then changing frequency of AE in one shot, P, can be calculated by:

$$P = \frac{N_{AP}}{N_{Total}} \quad (5)$$

Then, as instructed by cognitive informatics, audio pace (AP) can be formulated as follows:

$$AP = -1/Log(P) \quad (6)$$

which means that more frequently the audio energy changes, quicker audio pace is. It is perceptive that higher Audio Pace indicates tenser atmosphere or splendid scene. Therefore, viewer may pay more attention on the shot with high audio pace.

Same to the analysis in Section 3.2.1, audio energy and audio pace are not additive. Because they are all positive to auditory information, we integrate them as follows to formulate the auditory information, AI, by regarding audio energy as the weight of audio pace:

$$AI = AE * AP \quad (7)$$

3.2.3 Human perception descriptor

Because both visual and auditory information have the same unit, they are additive. By integrating both elements, we formulate human perception descriptor as follows:

$$HP(n) = \frac{\phi(VI(n) - \mu_{VI})}{\sigma_{VI}} + \frac{\psi(AI(n) - \mu_{AI})}{\sigma_{AI}} \quad (8)$$

where HP(n) denotes the human perception descriptor of n^{th} shot, μ and σ respectively denotes the mean and standard deviation of these features. Here, the weight ϕ and ψ are also initially set with 0.5, which means that both modalities, visual and audio modalities have the same influence on human perception.

3.3. Tempo formulation

Considering both aspects, we integrate elements form both film grammar and human perception for tempo computation as follows:

$$Tempo(n) = \gamma * FG(n) + \lambda * HP(n) \quad (9)$$

where Tempo(n) denotes the tempo of n^{th} shot. Then, with the variety of n in temporal domain, we will get tempo flow plot to depict the development of the storyline.

Here, the weight γ and λ are also initially set with 0.5, which means that film grammar, the external rule, and human perception, the internal feeling, contribute equally to movie tempo. Although this assumption reflects the relationship of both aspects to some extent, advanced analysis based on filmmaking and psychology is needed to improve tempo formulation in the future.

4. System Framework

We build an interactive system for action movie analysis and edit based on our work in [20]. The system consists of three steps: Structuralization, Action scene detection and Interactive interface. We will detailedly illustrate the three sections as follows.

4.1. Structuralization

Structuralization for movie content analysis includes two steps: shot boundary detection & keyframe extraction, and scene boundary detection.

4.1.1 Shot boundary detection& keyframe extraction

We detect the shot boundary and segment the video into shots by calculating the similarity of visual content between adjacent frames. Here, a 16*8 2D HS color histogram in the HSV color space of a frame is selected as the visual content. Then an unsupervised clustering based approach in [18] is used to extract key frames within a shot.

4.1.2 Scene boundary detection

Referring to [23], the problem of clustering shots into scenes is transformed into a graph partitioning problem. A weighted undirected graph called a shot similarity graph (SSG) is constructed to achieve the task. The SSG is then split into sub graphs by applying the normalized cuts for graph partitioning. The rule of partition is based on maximizing intra-subgraph similarities and minimizing inter-subgraph similarities. Besides, scene key frames are selected to describe the content of each scene.

4.2. Action scene detection

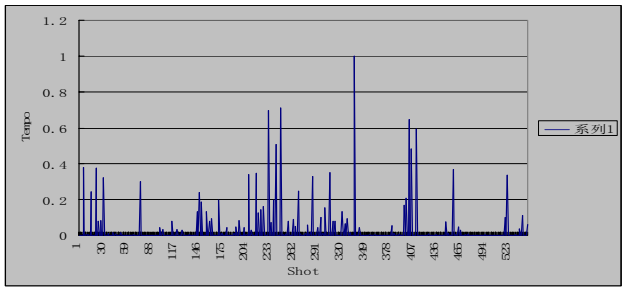
After calculation of tempo model, we get the tempo flow plot to depict the development of the storyline shown in Figure.1 (a). Moreover, the tempo flow plot is smoothed with a Gaussian filter with a size 9 window (σ =1.5), shown in Figure.1 (b), for the two reasons. Firstly, there will not be drastic changes in the continuously developing storyline. Secondly, human perception gradually changes because of memory retention.

The action scene is detected with the following rule:

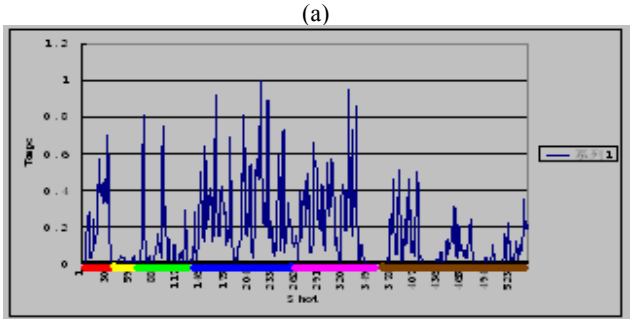
(a) Counting the number NP of the peaks within one scene, the values of which are higher than Threshold1;

(b) If NP is larger than Threshold2, the scene is considered as an action scene.

Here, Threshold1 and Threshold2 are empirical values derived from the experiments. From Figure.1 (b) and Figure.2, it is perceivable that because Scene 4 and 5 are action scenes, the tempo changes more drastically. With appropriate Threshold1 and Threshold2, action scene can be detected.



(a)



(b)

Figure. 1 (a) Tempo flow plot; (b) Corresponding tempo flow plot smoothed with Gaussian filter for one video clip of the movie, "Fearless" (Different colors denotes different scenes corresponding to Scene1-6 shown in Figure. 2).

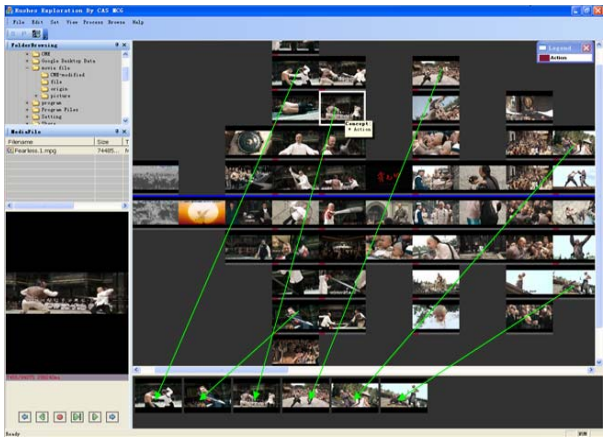


Figure. 2 Interactive interface of our system: hierarchical browsing of action movie, "Fearless".

4.3. System overview

As shown in Fig.2, our interface is composed of five parts: a folder browsing subwindow and a media file subwindow used for users to locate and select the movies they want to process; a playing-back subwindow (left-bottom) used for playing back the scenes or the whole video; a hierarchical browsing subwindow used for visualizing both the structural and semantic information; and finally a storyboard subwindow (right-bottom) used for users to select and reorder the interested shots or clips.

On the basis of structuralization, the video is segmented into scenes and key frames of each scene are extracted. For convenient navigation of movies, we put the video into a two-dimensional Cartesian Coordinates in the hierarchical browsing subwindow. Along the vertical dimension are the key frames of the same scene while along the horizontal dimension is the linear temporal dimension of scene sequences. Through this subwindow, editors/viewers can browse the structure of video briefly. Moreover, if editors/viewers want to know the specific contents of a certain clip, they can double click the corresponding key frame to launch a video player for playing back it in the left-bottom playing-back subwindow.

To annotate the action scene, we use a concept legend and a color bar under each key frame to display whether the corresponding scene is an action one. If an action scene is detected, all of its key frames are labeled with the same color. Furthermore, for convenience, the concept about a scene can also be displayed as mouse moving to the corresponding key frames.

Video editors/viewers can select their desired clips by dragging and dropping the corresponding key frames to the right-bottom storyboard subwindow. The clips in the box will be connected together from left to right to form a new video clip. Thus, editors/viewers can browse the movies and make movie summarization or trailer as they like.

5. Experiment Results

To demonstrate the effectiveness and generality of action scene detection with movie tempo, we select ten action movies depending on the classification in [19]. The details of each movie will be presented in Table 1. The previous test data only focus on "Action + Drama" and "Action + Sci-fi" movies. The characteristic of them is that their action scenes always depict one-on-one fighting. In our experiments, we find that tempo is a general feature and can be used for other kind of action movie. Therefore, we also selected four "Action + War" movies for test, the action scenes of which depicts the fierce battle with explosion, gunfire and so on forming the intense atmosphere, namely, the high tempo.

By human judgment, we found the ground truth for experimental result analysis. Precision and Recall are used to measure the results of action scene detection, which are well known rules in information retrieval field. The detailed experimental results are shown in Table 2.

From Table 2, we can see that recall value is satisfactory, which means that tempo can well depict the characteristics of action scene, including three kinds of action movies. Comparatively, precision value is a little lower. The main reasons are as follows:

(a). Some scenes, for example, the congested crowd walking in the street, have the most characteristics of action scene and thereby give the false positives.

(b). Although some scenes, for example, the scene with loud shout or noise, only have one or two features of action scene, the value is much bigger. The inappropriate parameters mentioned in Section 3 can not represent the best combination of the elements and lead to false positives.

6. Conclusion and future work

By analyzing film grammar and human perception, we present an innovative model of tempo and implement it on action scene detection for movie analysis. For the first time, we clearly propose that tempo indicates the rhythm of both movie scenarios and human perception. By thoroughly analyzing both aspects, we classify the elements of tempo into two sorts. The first is based on the film grammar as the previous research. The second is based on the human perception and we originally propose the information measure for perception depending on the cognitive informatics to describe the viewers' emotional changes to continuously developing storyline. With both aspects, tempo is defined and tempo flow plot is derived as the clue of storyline. On the basis of video structuralization and movie tempo analysis, we build a system for hierarchical browse and edit with action scene annotation.

Large-scale experiments demonstrate the effectiveness and generality of tempo for action movie analysis. However, the parameters for tempo computation mentioned in Section 3 should be deeply analyzed for better reflecting the relationship and importance of these elements. Besides, we will advance our research on analyzing other classifications of movies with tempo.

7. Acknowledgments

This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60773056) and Key project supported by Natural Science Foundation of Tianjin (No. 07JCZDJC05800).

References

[1] Howhard D. Wactlar. The Challenges of Continuous Capture, Contemporaneous Analysis and Customized Summarization of Video Content, CMU, USA.

[2] N.Vasconcelos and A.Lippman. Towards Semantically Meaningful Feature Spaces for the Characterization of Video Content. IEEE ICIP, 1997.

[3] Jeho, Nam et al. Audio-visual Content-based Violent Scene Characterization. IEEE ICIP, 1998.
[4] Brett Adams, Chitra Dorai, Svetha Venkatesh. Novel Apporach to Determining Tempo and Dramatic Story Sections in Motion Pictures. IEEE ICIP, 2000.
[5] Brett Adams, Chitra Dorai, Svetha Venkatesh. Study of Shot Length and Motion as Contributing Factors to Movie Tempo. ACM Multimedia 2000.
[6] Brett Adams, Chitra Dorai, Svetha Venkatesh. Role of Shot Length in Characterizing Tempo and Dramatic Story Sections in Motion Pictures. Proceeding of IEEE Pacific Rim Conference on Multimedia, 2000.
[7] Brett Adams, Chitra Dorai, Svetha Venkatesh. Toward Automatic Extraction of Expressice Elements from Motion Pictures: Tempo. IEEE Transactions on Multimedia, 2002.
[8] Lei Chen, Sariq J. Rizvi, M.Tamer Ozsu. Incorporating Audio Cues into Dialog and Action Scene Extraction. Proc. of SPIE Storage and Retrieval for Media Databases, 2003
[9] Alan F.Smeaton, Bart Lehane et al. Automatically Selecting Shots for Action Movie Trailers. Proceedings of the 8th ACM international workshop on Multimedia information retrieval, 2006.
[10] Hsuan-Wei Chen, Jin-Hau Kuo, Wei-Ta Chu, et al. Action Movies Segmentation and Summarization Based on Tempo Analysis. Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, 2004.
[11] Junyong You, Guizhong Liu, Li Sun, et al, A Multiple Visual Models Based Perceptive Analysis Framework for Multilevel Video Summarization, IEEE Transactions on Circuits and Systems for Video Technology, 2007.
[12] Yingxu Wang, On Cognitive Informatics, Proc. of the First IEEE International Conference on Cognitive Informatics, 2002.
[13] Yingxu Wang, On a New Frontier: Cognitive Informatics, Invited Talk, Proc. of the 7th International Conference on Object-Oriented Information Systems, Canada, 2001.
[14] D. Arijon. Grammar of the Film Language. Los Angeles, CA: Silman James Press, 1976.
[15] Leo Braudy and Marshall Cohen. Film Theory and Criticism: Introductory Readings. Oxford University Press, 1999.
[16] Hong-Wen Kang, et al. Space-Time Video Montage. Proceedings of International Conference on Computer Vision and Pattern Recognition, 2006.
[17] Bai Liang; Hu Yaali, Feature analysis and extraction for audio automatic classification, Proc. of IEEE International Conference on Systems, Man and Cybernetics, vol.1, pp:767-772,2005.
[18] Yueting Zhuang, Yong Rui, Thomas S. Huang et al. Adaptive key frame extraction using unsupervised clustering. Image Processing, ICIP 1998.
[19] http://www.apple.com/trailer.
[20] Sheng Tang, Yong-Dong Zhang, Jin-Tao Li et al. Rushes Exploitation 2006 By CAS MCG. In Proc. TRECVID Workshop,Gaithersburg,USA,Nov.2006..
[21] Shannon, C.E., A Mathematical Theory of Communication, ell System Technical Journal, 1948.
[22] Zeeshan Rasheed, Mubarak Shah. Detection and Representation of Scenes in Videos. IEEE Transaction on Multimedia, Vol7, NO.6, December, 2005.

Table 1. Detailed information about each movie

Movie Title	Fearless	Crouching Tiger, Hidden Dragon	Fist of Legend	Gladiator	The Matrix1	Minority Report	Enemy At The Gates	Wind talkers	Pearl Harbor	Thin Red Line
Runtime (min)	110	120	103	155	113	145	131	134	183	170
Movie Genre	Action+Drama				Action+Sci-fi		Action+War			
File Format	MPEG-1									
Audio Format	16 bits/sample, mono, 22kHz									
Delivery:(f/s)	25	30	25	30	30	30	30	25	30	25

Table 2. Experimental results

Movie Title	Fearless	Crouching Tiger, Hidden Dragon	Fist of Legend	Gladiator	The Matrix	Minority Report	Enemy At The Gates	Wind talkers s	Pearl Harbor	Thin Red Line
Ground	7	8	8	15	7	8	6	9	12	4
Detected	10	11	9	23	8	12	9	11	15	6
Falseaccepted	3	3	2	8	2	4	3	3	3	2
False rejected	0	0	1	0	1	0	0	1	0	0
Precision (%)	70	73	78	65	75	67	67	73	80	67
Recall (%)	100	100	88	100	86	100	100	89	100	100
Average Precision (%)	71									
Average Recall (%)	96									