SHOT CLASSIFICATION FOR ACTION MOVIES BASED ON MOTION CHARACTERISTICS

Shuhui Wang[1, 2, 3], *Shuqiang Jiang*[1, 2], *Qingming Huang*[1, 2, 3], *Wen Gao*[2, 4]
[1]Key Lab of Intell. Info. Process., Chinese Academy of Sciences, Beijing 100190, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[3]Graduate University of Chinese Academy of Sciences, Beijing, 100190, China
[4]Institute of Digital Media, Peking University, Beijing 100871, China
{shwang, sqjiang, qmhuang,wgao}@jdl.ac.cn

ABSTRACT

In this paper, we propose a shot classification method for action movies. Considering that motion characteristic is very important for semantic movie analysis, and it contains abundant information in action movies, the structure tensor analysis is used for feature extraction due to its capability of representing both spatial and temporal characteristics of a shot. Firstly, the movie shots with known labels are decomposed into a set of overlapped fixed-length segments and their structure tensor histogram are computed. The labels of segments are identical to the shots they belong to. Then *Adaboost* is used to train the semantic classifier with these structure tensor histogram sets. In testing procedure, the unknown shot are decomposed in the same way, and feature vector of each segment is extracted and classified by the classifier. Finally, the label of the shot is generalized by the segment label voting scheme. Experimental results show that this scheme could effectively deal with multiple motion patterns within shots and promising results are achieved.[1]

***Index Terms*—**Action Movies, Shot classification, Structure tensor feature, *Adaboost*, and Voting Scheme

1. INTRODUCTION AND OVERVIEW

Automatic content based movie understanding and indexing appears as an interesting issue due to the flourishing movie industry and users' need of efficiently browsing the movie database. Like any other video types, movie videos could be represented by hierarchical structure such as key frame, shot, scene, etc. Visual features like color, motion, texture could also be extracted for representing the raw data [1, 2, 3]. Shot identification and classification is the first step of semantic video indexing. There has been a lot of research in news videos and sports videos. In Ide *et al*'s work [4], five types of typical shots in news video are analyzed. Visual ontology like face is detected by template matching. Information from audio and text are utilized. A unified framework for sports video shot classification is proposed by Duan *et al* [5]. Eight domain irrelevant kinds of shot types are defined. Motion profile is computed from each motion vector field by mean shift clustering. After shot classification, domain knowledge is combined to conclude a semantic label.

Unlike sports video and news video which only combine limited environmental settings and several typical shots, shot types in movie are more diversified over various stories and genres. The semantic vocabulary of movies is open and could not be generalized by a finite set. This characteristic leads to the difficulty of automatic movie modeling and understanding. Currently, the mid-level concept detection for movie analysis is far from robust. Only reliable detection of some concepts has been used for high level semantic fusion and modeling. In Li and Weng's work [1] [2], they use face detection to detect a people, and face or speech recognition techniques to identify roles in the movie. The violence scene is characterized by combining discovery of explosion, blood, and the sound of gun-fire in Nam *et al*'s work [3]. However, these concepts are not enough for mining more information and high level semantic modeling.

Action movie is an important genre that attracts many audiences. It usually contains contents with strong motion that seems more likely to invoke people's attention and emotion [6]. These contents usually include shots with drastic motion such as running and chasing, men fighting, explosion and crash, etc. Identification of them will provide more information for further understanding of action movies.

In this paper, we present a motion based action movie shot classification method. Four kinds of shots are extracted and analyzed with each usually corresponding to distinct camera and local motion pattern. The feature we use is structure tensor histogram, which involves both neighboring pixel difference and all the pixels of the frames within a fixed-length segment, capable of representing the temporal and spatial motion characteristic of specific image sequence.

[1] This work was supported in part by National Natural Science Foundation of China under Grant 60773136 and 60702035, by National Key Technologies R&D Program under Grant 2006BAH02A24-2, and by National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117 and 2006AA01Z315

Even when there are multiple motion semantics in a shot, the corresponding segment in each time stamp has a unique semantic label.

Fig.1 is the overview of the proposed shot classification procedure. Either in training or testing step, all movie shots are split into many overlapped fixed length segments in the same way. Then they are converted into gray level images. Spatial-temporal reorganization is conducted, reorganizing the 3D shot volume into horizontal part *H* and vertical part *V*, and they are smoothed by a 3*3 low pass Gaussian filter. Structure tensor histogram is computed for each segment, thus a histogram set is formed for each movie shot. In training step, shots with unique semantic labels are split into segments and the corresponding features have the same label as the shot. *Adaboost* is used to train the classifier using labeled feature of segments. In testing step, the unknown shot are decomposed into segments as well. The classifier will generate a semantic label of the segments. Thus a set of semantic labels are collected for this shot. A fusion scheme is implemented to deduce a unique label for the shot. This scheme is capable of recognizing shots with multiple motion patterns and the effectiveness will be demonstrated by experiment.

This paper is organized as follows: In Section 2, we discuss the characteristics of four different shots in action movie. Structure tensor feature extraction is introduced in Section 3. Experiment results and discussion are presented in Section 4. We make a brief conclusion in Section 5.

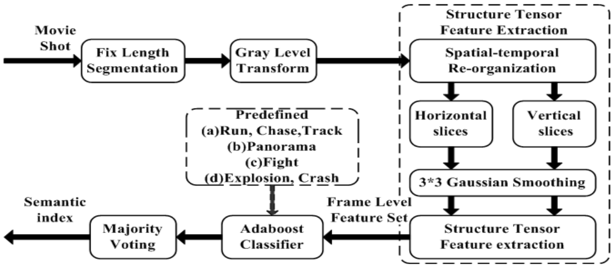

Fig 1. The overview of action movie shot classification

2. SEMANTIC CLASS DEFINITION

In this paper, the following four different semantic classes are defined and analyzed. (a)***Running, Chasing, and Object Tracking (RC)***: Strong camera motion like Pan or Tilt. (b)***Panorama (PAN)***: Slight pan, tilt and zoom, as well as low local motion intensity. (c)***Fighting (FI)***: Wavy camera motion patterns and irregular local motion patterns are usually involved. (d)***Explosion and Crashing (EC)***: No obvious camera motion and local image patches spread all over the whole frame. Fig 2 illustrates typical shots of these classes, where (a), (b), (c) and (d) represent ***RC***, ***PAN***, ***FI*** and ***EC*** respectively. The reason for selecting these classes could be stated from three aspects. Firstly, the dialogue shots could be identified with face and speech detection procedure [1, 2], so they are not included in this work. Secondly, although the four chosen semantic classes is still not a complete semantic set, it extends the above mentioned mid-level concepts with new elements like running, tracking and panorama. Finally, the motion patterns in these four classes are distinctive while they are quite ambiguous in other shot classes by our observation. Other features and knowledge could be utilized to identify more semantic classes, but this is beyond our consideration here.

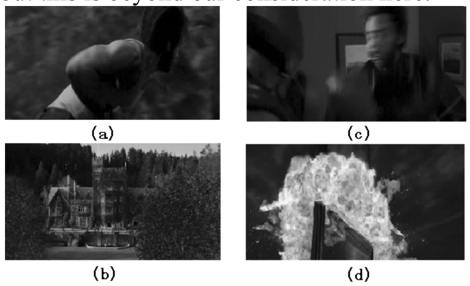

Fig 2. Four semantic classes: (a) ***RC*** (b) ***PAN*** (c) ***FI*** (d) ***EC***

3. MOTION FEATURE EXTRACTION

Motion analysis has been a popular issue for years since the rise of research in content based multimedia analysis. In Wang's work [7], a four parameter regression model is used to model the camera motion. An iteration procedure is built up and a frame mask is updated after each recalculation of parameters to discard the macro blocks with most different motion vector from the estimated global camera motion. Local motion is computed after camera motion estimation. However, it is very hard to precisely separate local motion from camera motion, because the regression model is easily biased by outliers, or even deviates in worse situations.

A non-parametric motion feature analysis is conducted by Duan *et al* [5], where motion profile is obtained by mean shift clustering on the P-frame motion vector field. Feature vectors are extracted based on motion profile, and the feature of a shot is generated by averaging P-frame feature set. However, there is usually one kind of global motion pattern in sports shots (e.g. Pan or Zoom), while there maybe more than two in action movie shots. Moreover, this scheme could not discover the temporal motion information. It may be vulnerable when processing shots with irregular motions. To overcome the above mentioned disadvantages, structure tensor analysis is used for feature extraction.

3.1. Structure Tensor Histogram

The characteristic of spatial-temporal pattern was discussed in detail by Ngo *et al* [8]. This method treats a video segments as a three dimensional volume considering all frames at a time, where three axes are width, height and frame index. As described in [8], the volume is decomposed into a set of 2-D temporal slices $H(x,t)$ and $V(y,t)$, where each is defined by plane (x, t) and (y, t) for horizontal and vertical slices, respectively. Pixel located in H is written as $H(x,t)|_{y=i}$, and similarly as $V(y,t)|_{x=j}$ in V. Take H as an example, the local structure of a slice $H|_{y=i}$ is consequently represented by tensor Γ as:

$$\Gamma = \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} \tag{1}$$

$$J_{xx} = \sum_{x',t' \in w} \widehat{H}_x^2 (x-x', t-t'), \widehat{H}_x = \partial(G*H)/\partial x$$

$$J_{tt} = \sum_{x',t' \in w} \widehat{H}_t^2 (x-x', t-t'), \widehat{H}_t = \partial(G*H)/\partial t$$

$$J_{xt} = \sum_{x',t' \in w} \widehat{H}_x (x-x', t-t') \widehat{H}_t (x-x', t-t')$$

Here G is the Gaussian smoothing kernel and w is the 3*3 support window centered at each pixel in a slice.

The rotation angle θ indicates the direction of gray level change in support window w. It could be written as:

$$R\begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} R^T = \begin{bmatrix} \lambda_x & 0 \\ 0 & \lambda_t \end{bmatrix}, R = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{2}$$

We have local orientation ϕ of the local structure w:

$$\phi = \begin{cases} \theta - \frac{\pi}{2}, \theta > 0 \\ \theta + \frac{\pi}{2}, else \end{cases}, \phi \in [-\frac{\pi}{2}, \frac{\pi}{2}], \theta = \frac{1}{2}\tan^{-1}\frac{2J_{xt}}{J_{tt} - J_{xx}} \tag{3}$$

A certainty measure c is computed as:

$$c = \frac{(J_{xx} - J_{tt})^2 + 4J_{xt}^2}{(J_{xx} + J_{tt})^2}, c \subset [0,1] \tag{4}$$

After figuring out ϕ and c, the structure tensor histogram could be represented by:

$$M(\overline{\phi}, t) = \begin{cases} \sum_i \sum_x c(x,t)_{|y=i}, if\, \phi(x,t)_{|y=i} \subset \overline{\phi} \\ 0, else \end{cases} \tag{5}$$

where $\hat{\phi}$ is a nonuniformly quantized bin. The number of bins is 9 in our experiment, and the boundaries are set by 5-folds cross validation: $[\frac{-\pi}{2}, \frac{-\pi}{3}, \frac{-\pi}{4}, \frac{-5\pi}{36}, \frac{-\pi}{36}, \frac{\pi}{36}, \frac{5\pi}{36}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}]$.For V slices, the same computing procedure is conducted.

3.2. Feature Computation of Segments

We can easily find that the computation of each structure tensor histogram $M(\overline{\phi}, t)$ involves 5 frames considering the 3*3 smoothing window and the double-sided difference operator $\partial A / \partial t$, where A stands for the gray level representation of the whole image. The shot is firstly divided into several overlapped segments, and the length of each segment is 13 frames. For instance, in the first segment, 1st, 4th, 7th, 10th and 13th frame is extracted to build a 5 frame sample set. The next segment contains frames of 4th to 13th in the first segment, and another part from 14th to 16th frame, and 4th, 7th, 10th, 13th, and 16th frame are regrouped to form another 5-frame sample set. This procedure repeats until the end of the shot. For each shot, the number of segments is:

Num_Seg = ceil ((shotend – shotbeg) / 3) - 3 (6)

Here *shotend* and *shotbeg* represents the frame index of the end and the beginning of the shot. After computing the structure tensor histogram for each segment, feature normalization is operated as:

M(i) = M(i) / SUM (M(i)), i = 1,…18.
M(19) = log(SUM (M(i))) (7)

The reason for using log for computing M(19) is that it will not change the monotony property of *SUM (M(i))*, and the magnitude could be reduced since the value is very large. Therefore, a 19 dimensional feature vector is computed by Eq.(7) for representing a segment within a shot.

3.3. The Voting Scheme

In testing procedure, we use the result of segments voting to identify the unknown shot. The scheme could be formulated as: $shot\_idx = arg_i\, max(S(i))$. Here S is a histogram with four bins corresponding to the four classes. It identifies the shot by the label with the most segments classified as.

4. EXPERIMENT

4.1. Data Collection and Manual Annotation

In this paper, five action movies have been collected for shot classification experiment, which are: *The Matrix*, *The Matrix Reloaded*, *Kill Bill I*, *X-Men III*, and *Minority Report*. Though slightly different in resolution, the width and height are not adjusted for it may bring about distortion of objects in a frame. We choose the shots from all the 5 movies, with their shot boundary precisely manually marked.

There are a few shots with more than two semantic labels. We label them with the dominant in the shot. For other shots with disambiguity, we assign them with a unique semantic label. All the segments belonging to a shot is assigned with the same label. Programs are run on Intel Core2 desktop with 1.86GHZ CPU speed, and 1G Memory. Processing speed is about 2 *segs/sec*.

Table 1 The Constitution of the four semantic classes

	RC	PAN	FI	EC	Total
Segments	3158	3038	3961	1543	11700
%	27	25.97	33.85	13.19	100
Shots	150	33	257	119	559
%	26.83	5.90	45.97	21.30	100

4.2. Experiment on segment identification

All of the 559 shots are equally split into 5 subsets. Four are used for training and one for testing each time, and we choose different testing subset 5 times. It must be emphasized that the training elements are structure tensor histograms of segments, while *shot* means a collection of its own segments. After feature extraction, we format the data into *weka* [9] arff file, and use *Adaboost* [10] for training and classification. The weak classifier is *J48* decision trees. Iteration times in *Adaboost* are set to 30. Table 2 shows the result of segments identification. Table 3 is statistical measurement of Table 2, where *Precision (P)*, *Recall (R)* and *F-measure (F)* are evaluated.

Experimental result in Table 3 shows that the *Recall* of ***EC*** is lower than other classes. The reason may lie in the motion complexity of explosion and crash. The texture of flame is rather blurry. However, the result could be improved by employing other features like color features. The result of ***PAN*** is the best of all because shots of ***PAN*** usually contain slight camera motion and local motion. Thus ***PAN*** is well separated from others.

Table 2 Result on Segment identification

a	b	c	d	←Labeled as
2635	45	430	48	a = ***RC***
41	2591	33	13	b = ***PAN***
353	32	3466	110	c = ***FI***
159	28	326	1030	d = ***EC***

Table 3 Statistical Measure of Segments identification

<table>
<tr><td colspan="2"><i>Correctly</i></td><td>10082</td><td>86.171 %</td></tr>
<tr><td colspan="2"><i>Incorrectly</i></td><td>1618</td><td>13.829 %</td></tr>
<tr><td></td><td><i>P</i></td><td><i>R</i></td><td><i>F</i></td></tr>
<tr><td><b><i>RC</i></b></td><td>0.827</td><td>0.834</td><td>0.831</td></tr>
<tr><td><b><i>PAN</i></b></td><td>0.966</td><td>0.971</td><td>0.968</td></tr>
<tr><td><b><i>FI</i></b></td><td>0.815</td><td>0.875</td><td>0.844</td></tr>
<tr><td><b><i>EC</i></b></td><td>0.858</td><td>0.668</td><td>0.751</td></tr>
</table>

4.3. Semantic fusion for shot Label generation

Evaluation on shot label generation is conducted, equally splitting dataset the same way as in 4.2. The predicted labels of segments are used for semantic fusion.

Table 4 Semantic Fusion of Voting Scheme

a	b	c	d	←Labeled as
120	2	25	3	a = ***RC***
4	27	2	0	b = ***PAN***
8	1	243	5	c = ***FI***
24	3	38	54	d = ***EC***

Table 5 Feature Averaging

a	b	c	d	←Labeled as
58	1	72	19	a = ***RC***
2	16	8	7	b = ***PAN***
40	4	180	33	c = ***FI***
31	10	54	24	d = ***EC***

Table 6 Statistical Measurement

<table>
<tr><td>Classified</td><td colspan="3">Voting Scheme</td><td colspan="3">Feature Averaging</td></tr>
<tr><td><i>Correctly</i></td><td>444</td><td colspan="2">79.43%</td><td>278</td><td colspan="2">49.73%</td></tr>
<tr><td><i>Incorrectly</i></td><td>115</td><td colspan="2">20.57%</td><td>281</td><td colspan="2">50.27%</td></tr>
<tr><td></td><td colspan="3">Voting Scheme</td><td colspan="3">Feature Averaging</td></tr>
<tr><td></td><td><i>P (%)</i></td><td><i>R (%)</i></td><td><i>F</i></td><td><i>P (%)</i></td><td><i>R (%)</i></td><td><i>F</i></td></tr>
<tr><td><b><i>RC</i></b></td><td>76.92</td><td>80</td><td>0.784</td><td>44.27</td><td>38.67</td><td>0.413</td></tr>
<tr><td><b><i>PAN</i></b></td><td>81.82</td><td>81.82</td><td>0.818</td><td>51.61</td><td>48.48</td><td>0.500</td></tr>
<tr><td><b><i>FI</i></b></td><td>78.90</td><td>94.55</td><td>0.860</td><td>57.32</td><td>70.04</td><td>0.630</td></tr>
<tr><td><b><i>EC</i></b></td><td>87.10</td><td>45.38</td><td>0.597</td><td>20.17</td><td>28.92</td><td>0.238</td></tr>
</table>

Besides our voting scheme, frame feature averaging scheme [5] is used for comparison. Table 4 and 5 shows the result of voting schemes and feature averaging. Table 6 is statistical measurements. The result shows the advantage of our scheme. However, when the number of segments in a shot is very small, say, 1 or 2, the result is heavily influenced by the error of segment classifier. Though such shots are minority, this problem is very challenging.

5. CONCLUSION

In this paper, we propose a method to solve the problem of shot classification for action movies. Structure Tensor Histograms feature is extracted and *Adaboost* is used for generating semantic classifier by training using features of segments rather than shots. Experiments on both segment identification and shot label generation are conducted. Comparison between voting fusion and feature averaging is made. Our method provides promising result of shot classification for action movies. In future work, more kinds of shots will be collected. More semantic classes as well as their correlations will be considered. Meanwhile, we will be dedicated to improving the shot classification algorithm by better modeling of feature space and semantic space.

6. ACKNOWLEDGEMENT

The authors would thank the support by "Science100 Plan" of Chinese Academy of Sciences under Grant 99T3002T03.

7. REFERENCE

[1] Y. Li, S. Narayanan, C. C. J. Kuo, "Content-based Movie Analy-sis and Indexing Based on AudioVisual Cues", in *IEEE Trans. CSVT*, vol. 14, no. 8, pp. 1073-1085, 2004.

[2] C. Y. Weng, W. T. Chu, J. L. Wu, "RoleNet: Treat a Movie as A Small Society", in *Proc. ACM Int. Conf. MIR*, 2007, pp. 51-60.

[3] J. Nam, M. Alghoniemy, A. H. Tewfik, "Audio-Visual Content Based Violent Scene Characterization", in *Proc. IEEE. Int. Conf. Image Process*., 1998, pp. 353-357.

[4] I. Ide, K. Yamamto, H. Tanaka, "Automatic Video Indexing Based on Shot Classification", in *Conf. Advanced Multimedia Content Processing*, 1998, vol. 1554, pp. 87-102.

[5] L. Y. Duan, M. Xu, Q. Tian, C. S. Xu, J. S. Jin, "A Unified Framework for Semantic Shot Classification in Sport Video", in *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1066-1083, 2005.

[6] H. W. Chen, J. H. Kuo, W. T. Chu, J, L. Wu, "Movies Segmentation and Summarization Based on Tempo Analysis", in *Proc. ACM MIR*, 2004, pp. 251-258.

[7] R. Wang, T. S. Huang, "Fast Camera Motion Analysis in MPEG Domain", in *Proc. IEEE. Int. Conf. Image Process.*, 1999,vol. 3, pp. 691-694.

[8] C. W. Ngo, T. C. Pong, H. J. Zhang, "Motion Analysis and Segmentation through Spatial-Temporal Slices", in *IEEE Trans. Image Processing*, vol. 12, no. 3, pp. 341-354, 2003.

[9] *Weka 3*: website: http://www.cs.waikato.ac.nz/ml/weka/

[10] FreundY, Schapire RE, "A short introduction to boosting", *J Japan Soc Artif Intell* 14(5):771－780, 1999.