http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (1): 98-105, 2011
ISSN 1812-5638

Semantic Scene Segmentation for Advanced Story Retrieval

Songhao Zhu and Zhiwei Liang
School of Automation, Nanjing University of Post and Telecommunications,
9, Wen Yuan Road, Xi Xia Distinct, Nan Jing, 210046, China

Abstract: In this study, the issue of categorizing video scenes into semantic classifications is addressed with respect to the information of audio-visual cues. More specifically, the grammar of film production is first exploited to specify scene contents with respect to human perception. Next, each scene is categorized into one of the following three classes: conversation, action and suspense. To achieve more specific scene and consist with human perception, conversation scene are further categorized into emotional conversation and common one and action scene are categorized into gunfight, beating and chasing scene. This study is a step toward browsing and retrieval of feature films on the internet in limited bandwidth, video repository and rating of feature films of interest.

Key words: Video feature, audio feature, film grammar, human perception, effectively, efficiently

INTRODUCTION

Nowadays, feature film collections is becoming available to more and more average consumers, accompanied with cheaper storage devices, higher transmission rates and advanced compression techniques. The demand for efficient tools to retrieve contents of interest from these large-scale video databases is increasing tremendously. That is, categorizing video scenes into conceptually semantic scenes to cater to perceptual user studies is becoming a hot research in the field of multimedia application.

Many approaches have been developed for scene categorization, which can be classified into two genres: for movie videos and for special videos. Yoshitaka *et al.* (1997) adopted shot duration time and visual dynamic activities to differentiate different scene classifications. In their work, shot visual dynamics is computed based on the color statistics of frames in the shot and the repeating shots are clustered into the same scene according to their similarities. Adams *et al.* (2000) categorized scene genres according to the tempo in movies, which is formulated as a continuous function based on two features: camera motion magnitude and shot length. Sundaram and Chang (2000) cluster visually similar shots into the one and only dialog scene using the information of Eigenface. Similarly, Pfeiffer *et al.* (2001) utilized the information of human face to group the similar shots into the same scene. Shots showing the same actor with similar position and size are first clustered into the same face-based class. Then, these face-based classes are further linked across shots to form the face-based sets based on the Eigenface information. Finally, the pattern of a dialog scene can be labeled if several conditions are simultaneously satisfied. Li *et al.* (2003) classify a scene into one of three scenes: two-speaker dialog, multi-speaker dialog and others using the structural information of a scene based on the color information of the shots in the scene. In their approach face information, which is an important and significant cue for speaker detection, is not used. Tavanapong and Zhou (2004) use gray information in local region to analyze respective characteristic of different scene categories and the defined scene categories are not so specific: traveling, serial-event and parallel-event. Zhao *et al.* (2007) parse video scene into each of following three types: parallel scene with interesting event, parallel scene with simultaneous serial event and serial scene, which is not so specific and not consist with human perception. These methods are based on the grammar of film, which is a set of production rules of how the movies should be composed. Compared with their approach, our framework analyzes the structure of movie scenes, uses the features of low-level and middle-level and classifies scenes into more specific categories.

There is a particular interest in the scene categorization of the special videos, such as news video and sports video. Color information is utilized to achieve the classification of soccer scenes (Xie *et al.*, 2004). Huang *et al.* (2005) deal with the problem of scene categorization using Hidden Markov Model. These approaches are not available in other domains of videos, such as feature films. Furthermore, the scene classification

Corresponding Author: Songhao Zhu, School of Automation, Nanjing University of Post and Telecommunications, 9, Wen Yuan Road, Xi Xia Distinct, Nan Jing, 210046, China Tel: +86-25-85866500, +86-15952021713

of special videos often involves the special treatment, such as the anchor in news videos and shooting in sports videos.

In this study, we propose a scheme for categorizing video scenes into specific scene genres: conversation scene including emotional and common categories, action scene containing gunfight, beating and chasing classes and suspense scene based on the grammar of film. This approach exploits the structural information of predefined specific scenes based on the information of both audio and visual which are robust and easily computable. We believe that low-level (such as color and audio energy) and middle-level features (such as human face) may be combined with the high-level domain knowledge and which helps to improve the accuracy of scene classification.

Compared with these above state-of-the-art algorithms, the proposed scheme has following two advantages. On the one hand, the proposed approach categorizes scene into more specific classes, which can be more suitable for human understanding. On the other hand, our approach utilizes the characteristics of movie scenes and the grammar of film editing to address the problem of scene classification for high-level retrieval.

PROPOSED SCHEME

Movie directors often utilize certain rules to specify the genre of a scene. Such rules are the so-called film grammar in the film literature (Bordwell and Thompson, 1997). By following such film grammar, camera movements, sound effects and lighting distributions can convey the information of mood and atmosphere to viewers. Although, different directors may use film grammar differently, scenes of the same genre have many common features. For example, most of the action scenes have similar tempo and sound. Our aim is to analyze these audio-visual cues to determine the classification of one given scene.

We first describe the characteristics of three conceptually meaningful scenes. Then, we classify scenes into one of three categories: conversation, action and suspense by making use of the low-level features including audio-visual information and mid-level shot feature containing human face information. Finally, we make three subclasses; gunfight, beating and chasing under action group and categorize conversation scenes into emotional category and common category. Figure 1 illustrates the whole process of the classification of scene genres. This is done by analyzing the information of color and audio and combining that with the film grammar to classify scenes.

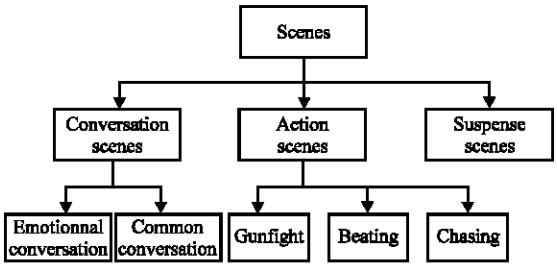

Fig. 1: Flow chart presenting the classes of scene genre

Shot detection and scene identification: We use the algorithm (Zhu and Liu, 2009a, b) for the detection of shot boundaries using the two-dimension histogram entropy intersection. Let D(k) represents the intersection of histograms entropy H^k and H^{k-1} of frames k and k-1 respectively. That is:

$$D(k) = \sum_i \sum_j \min\left(E_{ij}^k, E_{ij}^{k-1}\right) \quad (1)$$

Then we define the shot change measure T(k) and T(k+1) as:

$$\begin{cases} T(k) = D(k) - D(k-1) \\ T(k+1) = D(k+1) - D(k) \end{cases} \quad (2)$$

Shot boundaries are detected by setting a threshold on T:

$$\begin{cases} T(k) < -T_{change} \\ T(k+1) > T_{change} \end{cases} \quad (3)$$

Representing the content of a video shot concisely is a necessary step for various video processing. In this study, we extract key frames for each shot based on the dynamics of visual content using the algorithm proposed by Zhu and Liu (2009a).

To avoid under segmentation or over-segmentation of scenes, we use the spatio-temporal correlation to identify scene as the algorithm presented by Zhu and Liu (2009a, b).

Representative characteristics of film scenes: Based on the grammar of film used in the continuously recorded video, different genres of scene have their correspondingly representative characteristics.

- **A conversational scene:** Faces with similar spatial position and similar size, a sequence of shots with low activity intensity and shots with similar visual content alternatively present

- **An action scene:** A number of temporally continuous shots with short duration time, intensive action activity and/or intensive audio energy
- **A suspense scene:** A group of shots with low illumination intensity, a long period of low audio energy and low activity intensity followed by a sudden change either in sound track or in activity intensity or both

Appropriate features should be chosen to depict scene content sufficiently. In the experiment, besides visual content such as human face information, illumination intensity information and activity intensity information, audio content such as audio energy information are also utilized to describing the typical characteristics of semantically meaningful scenes.

Audio features selection: In order to achieve the accuracy of the scene classification, audio information is an important and necessary clue. Because feature extraction is very important for audio content analysis, the extracted features should capture the temporal and spectral structure of different audio classes from each scene talked above. Here, following features are selected to complete the task of audio clip classification, such as zero-crossing rate, energy envelope, spectrum flux, band periodicity, mel-frequency cepstral coefficients, spectral power and linear prediction based cepstral coefficients.

Furthermore, in this experiment, all audio clips are divided into non-overlapping sub-clips. A sub-clip is of one second duration and is further divided into forty twenty-five millisecond-long frames and short-time energy envelope entropy. The classification is performed based on these one-second sub-clips.

- **Zero-crossing rate:** Zero-crossing rate is a simple measure of the audio signal frequency content. The N-length short-time zero-crossing rate of the nth audio frame s(n) is defined as:

$$ZCR(n)=\frac{1}{N}*\sum_{t=n-N+1}^{n}\frac{|\,sgn\{s(t)\}-sgn\{s(t-1)\}\,|}{2}w(n-t) \tag{4}$$

where w(n) is a rectangular window and

$$sgn\left[s(n)\right]=\begin{cases}1 & s(n)\geq 0\\ -1 & s(n)<0\end{cases} \tag{5}$$

- **Energy envelope:** Energy envelope is used to calculate the global temporal information. It can be computed in following way: third order Butterworth low-pass filtering of the analytical signal root mean square amplitude of each audio frame:

$$EE(n)=\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left[s_t(n)\right]^2} \tag{6}$$

where N is the number of sample points in the nth audio frame

- **Spectrum flux:** Spectrum flux is defined as the two-norm of the frame-to-frame spectrum amplitude difference vector:

$$SF(n)=\left\|\,|M_f(n)|-|M_f(n+1)|\,\right\| \tag{7}$$

where, $|M_f(n)|$ is the magnitude of the FFT of the n^{th} frame at frequency value f. Both magnitude vectors are normalized in energy. Spectrum flux is a measure of spectral change between the adjacent two frames

- **Band periodicity:** Band periodicity describes the property of each sub-band. In this study, we consider four sub-bands including 500~1000 Hz, 1000~2000 Hz, 2000~3000 Hz and 3000~4000 Hz, respectively. The periodicity property can be represented by the maximum local peak of the normalized correlation function
- **Mel-frequency cepstral coefficients (MFCCs):** It has been proved that the mel-frequency cepstrum is a useful and highly effective feature in modeling the subjective pitch and frequency content of audio signals. The MFCCs are computed from fast Fourier transformation:

$$\begin{cases}MFCC(n)=\sqrt{\frac{2}{J}}\sum_{j=1}^{J}\left\{\left(\log S_j\right)\times\cos\left[t(j-0.5)\frac{\pi}{J}\right]\right\}\\ t=1,2,.....,T\end{cases} \tag{8}$$

where, the parameter J denotes the number of band-pass filters and the parameter T means the order of the cepstrum. In our scheme, J and T are set to be 24 and 12, respectively, namely the 24 band-pass filters and 12-order MFCCs are used

- **Spectral power:** Spectral power of each audio frame is computed with a Hanning window h(n):

$$h(n)=\sqrt{\frac{2}{3}}\times[1-\cos(2\pi\frac{n}{N})] \tag{9}$$

The spectral power of the nth audio frame s(n) is:

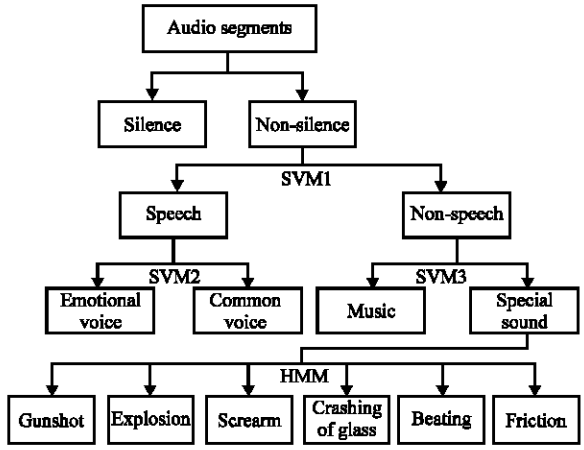

Fig. 2: Audio classification scheme

$$SP(k) = 10\log\left[\frac{1}{N}\left\|\sum_{n=0}^{N-1} s(n)H(n)\exp(-j2\pi\frac{nk}{N})\right\|^2\right] \quad (10)$$

- **Linear prediction based cepstral coefficients (LPCCs):** LPCCs are utilized to represent the timbre information of the voice components. The basic idea behind linear predictive analysis is that an audio frame can be approximated as a linear combination of past audio frames. By minimizing the sum of the squared differences over a finite interval between the actual audio frames and the linear predictive ones, a unique set of predictor coefficients can be determined

Audio signal classification: After removing silent clips, audio clips are classified into two categories firstly, i.e., speech and non-speech, based on the information of zero-crossing rate, energy envelope and spectrum flux. Then, speech clips are classified into emotional voice and common voice based on the spectrum flux, band periodicity, mel-frequency cepstral coefficients and non-speech clips are classified into special sound and music based on the spectral power and Linear prediction based cepstral coefficients. Finally, special sound is further classified into several classes, including gunshot, explosion, scream, beating, crashing of glass and rubbing of tires using hidden markov models (HMM). Figure 2 illustrates the idea of the classification scheme. From this Fig. 2, we can clearly see the whole classification process of the audio information.

For support vector machines-based classification, features are first combined to construct a feature vector. Then, the mean and standard deviation of these feature vectors over all forty audio frames are calculated and these statistics compose another feature vector. Finally, this new feature vector is normalized by its standard deviation of training data. The normalized feature vector is considered as the final representation of one-second audio signal.

The information of timbre and rhythm is utilized in the generative model for recognition, namely hidden markov models. Timbre allows one to tell the difference between sounds at the same loudness made by different objects. Each kind of timbre is denoted by one state of HMM and represented with the Gaussian mixture density. Here, rhythm is adopted to represent the change pattern of timbres and is denoted by the parameters of transition and duration in HMM.

Visual features selection

- **Face Information:** The occurrence of face is a salient feature in video, as it means the present of human in the scene. The size of a face is also a hint for the role of the person, i.e., a large face denotes that this person is in the center of attention. In the experiment, we use the face detection algorithm proposed by Li *et al.* (2002) which performs reasonably good for faces with different scales in the video. As a result of the face feature extraction process, we obtain the position and size of each detected face and the number of hits
- **Activity intensity:** The activity intensity indicates the tempo in video. For example, in conversational scene, the activity intensity is relatively low. On the other hand, in action scene, the activity intensity is relatively high. The activity intensity of the kth scene is:

$$AI(k) = \frac{1}{N_k - 1}\sum_{l=1}^{N_k-1}\min\left[\sum_{i=0}^{i=L-1}\sum_{j=0}^{j=L-1}\left(H_{ij}^{l}, H_{ij}^{l+1}\right)\right] \quad (11)$$

Figure 3a and b show a set of diagrams of activity intensity for different types of scene

- **Illumination intensity:** Bordwell and Thompson (1997) indicates the amount and distribution of light in relation to shadow and darkness and the relative tonal value of the scene is a primary visual means of setting mood. The amount of light in scene Sc(k), here, is described as the average illumination intensity:

$$\begin{cases} II(k) = \dfrac{\sum_i ShAvgII(i) \times ShLen(i)}{N(k)} \\ ShAvgII(i) = \dfrac{\sum_j KfII(j)}{N(i)} \end{cases} \quad (12)$$

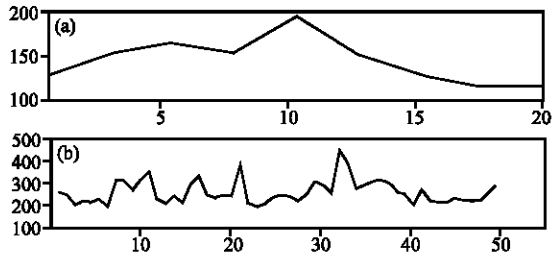

Fig. 3: A set of diagrams of activity intensity of (a) a TV interview, (b) the movie X Man III. The label of X is the number of scenes in video data and the label of Y is the activity intensity

where, ShAvgInt(i) is the average illumination intensity of the entire key frames in the ith shot, ShLen(i) is the length of the ith shot in terms of frames and N(k) is the total number of frames within the kth scene. Furthermore, KfInt(j) is the illumination intensity of the jth key frames in the ith shot and N(i) is the total number of frames within the ith shot

We use the average illumination intensity histogram to illustrate the average intensity distribution for three different illustrate the average intensity distribution for three different scenes in Fig. 4a-c.

- **Average duration:** Similar to activity intensity, average duration in terms of frames is another measurement of the video tempo and is computed as follows:

$$AD(k) = \frac{1}{N_k}\sum_{l=1}^{N_k} ShLen(l) \qquad (13)$$

Dialogue scene determination

- **Dialogue scene determination:** To locate the dialogue scenes, we first adopt the information of face to detect shot sequence with alternately recurring visual contents, which include similar size, similar position and same number. Figure 5 shows an example of dialogue-like scene.

Given these dialogue-like scenes, we exploit their corresponding audio information to further make sure they are actual conversation contents. Specifically, we should differentiate speech signals from music and other sounds. Here, a simple classification method is utilized to complete the task of discrimination using zero-crossing rate, energy envelope and spectrum flux as shown in Fig. 2.

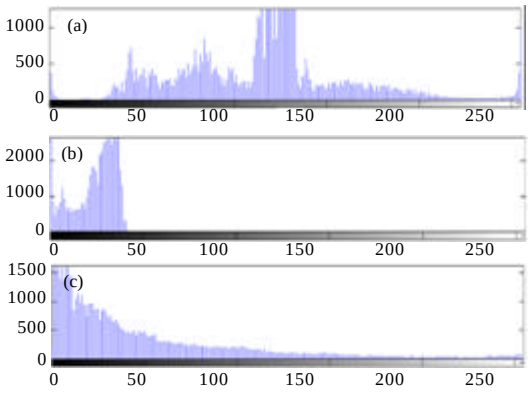

Fig. 4: Average intensity histogram of (a) a conversational scene in the movie Mission Impossible III; (b) a suspense scene in the movie The Ring; (c) an action scene in the movie X Man III. The label of X denotes the gray-level information and the label of Y describes the corresponding number of frames in each gray-level

Fig. 5: A dialogue-like scene detected using face information

- Emotional Dialogue Scene Determination. Among many dialogues, emotional conversations often attract viewers' attention and effect upon the flow of story. To discriminate the emotional contents from common ones, two acoustic features including average pitch frequency and temporal intensity variations are used. The first feature is estimated by 12-order linear prediction based cepstral coefficients and the second one is represented by the variance of spectral power levels over all forty signal segments within one-second audio sub-clip

Active scene discrimination

- **Active scene discrimination:** Similar to dialogue scene, active scene is another conceptually meaningful story content. Among active scenes, gunfight scene, beating scene and chasing scene are often the most interesting events and can instantly attract viewers' attention in films. Therefore, based on the recognition of active scene by integrating audio-visual signatures, three specific and distinct events are identified one-by-one

(a) (b) (c)

Fig. 6: The most typical visual features of gunfight scene including (a) gunfire; (b) explosion; (c) bleeding

According to the characteristics of active scene as talked above, these scenes with average scene duration less than twenty-five frames and average scene activity intensity is larger than two hundred or/both average audio energy is larger than one hundred are identified as active scenes.

Among active scenes, gunfight, beating and chasing events are three important types and mostly the climaxes in feature films. Next, we will detect these important active scenes using their unique audio-visual signatures.

- **Gunfight scene discrimination:** Gunfire, explosion and bleeding are the most typical visual features of gunfight scenes as shown in Fig. 6

Compared to gunfire cases, flames from an explosion show longer duration and cover wider areas. However, flames from the explosion and gunfire case both have dominant yellow, orange and/or red color histogram. Therefore, a predefined color table containing a certain range of color values is here adopted to discriminate the gunfight-like scenes.

Since, some violent actions, such as beating, gunshot and explosion can result in bleeding, bleeding is considered as another violence-related visual feature of gunfight event. We detect bloody color pixels using simple pixel-matching with the predefined color table.

Since, other events may have similar visual features as gunfire, explosion and bleeding, the audio information provides a supplement to the detection of gunfight scene. A distinct feature of gunfight scenes is the unique sound track. Specifically, given the audio track for successive gunfight-like shots, we discriminate its class based on a hidden markov model. The likelihood ratio between the input audio track and the defined sound classes is calculated to determine which class the associated sound belongs to as shown in Fig. 2.

- **beating or chasing scene identification:** In general, beating or chasing events (as shown in Fig. 7 and 8, respectively) are inherently accompanied by unique sound (e.g., beating, rubbing of tires, etc.)

Fig. 7: An example of chasing event

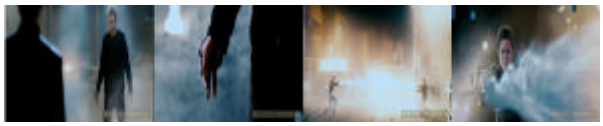

Fig. 8: An example of beating event

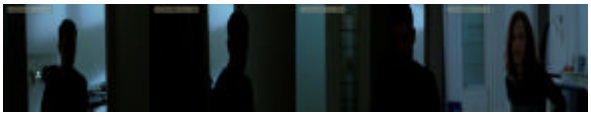

Fig. 9: An example of suspense scene

In particular, for active scene, we identify its specific class (beating or chasing) based on the likelihood ratio between its audio track and the given sound classes as shown in Fig. 2.

Suspense scene detection:

- **Suspense scene detection:** Suspense scenes are often the most events in horror and detective movies and instantly attract a viewer's attention in horror and detective genres. Figure 9 shows one example of suspense scene

According to the unique characteristics of suspense scene a scene can be declared as a suspense scene if following criterions are satisfied simultaneously:

- Average illumination intensity is less than 50
- There exist shots with audio energy envelope change suddenly from 5 to over 50
  Or / both

There exist shots with activity energy change instantly from 5 to over 100.

EXPERIMENTS

Here, we first analyze the experimental data and then evaluate the proposed scheme on various film genres.

Experimental design: We have tested the performance of the proposed scheme for classifying video scenes into

semantically meaningful categories over seven movies containing X Man III, Mission Impossible III, The Ring, Death On The Nile, The Sound Of Music, A Walk in the Clouds and The Girl Next Door. These feature films cover various genres such as action, horror, espionage, Musicals and comedy. All video sequences are digitized at 320×240 spatial resolutions with the rate of 15 frames/sec. Synchronized audio signals are sampled at 22 kHz and represented by 16-bits per sample.

To quantitatively evaluate the quality of semantic scene categorization, we invited 20 volunteer subjects, including 10 males and 10 females, to choose the most suitable category from three defined scenes for each video clip. Then, each video clip is labeled with a category that most of the 20 volunteer subjects agreed upon and the given label is considered as the corresponding ground truth label.

Similar to previous works on semantic scene classification, precision, recall and micro F1 are used as the performance measures (Yang and Liu, 1999). The formulation of precision, recall and micro F1 are described as follows:

$$\begin{cases} \text{Precision} = \dfrac{Nc}{Ns}, \quad \text{Recall} = \dfrac{Nc}{Ng} \\ F1 = \dfrac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \end{cases} \tag{14}$$

where, Nc denotes the number of detected scenes with correct semantic category, Ns denotes the number of detected scenes and Ng denotes the number of ground truth scenes with certain semantic category.

EXPERIMENTAL RESULTS

The overall performance of the semantic classification of scene content is summarized in Table 1. It can be observed that the average precision and recall are 0.924 and 0.920, respectively, which clearly demonstrate that the proposed scheme can classify video scenes into their inherent categories. From another point of view, the selected features for depicting scene content and the proposed way of categorizing scene content together achieve satisfactory experimental results.

To evaluate the effectiveness of the proposed approach, we compare it with the approach proposed by Yoshitaka *et al.* (1997) since it also defines specific scene genres. The results are shown in Table 2.

From Table 2, it can be seen that average F1-Score are 0.922 and 0.849, respectively. That is, the proposed approach exhibits a gain 8.60% of the average F1-Score compared with the method proposed by Yoshitaka *et al.* (1997). The reason why the proposed approach gains a large improvement in all the evaluating measures is based on following two phases. On the one hand, the low-level features and middle-level features are integrated into the procedure of the classification of scene genres. On the other hand, besides visual features, audio features are also utilized to help to improve the system performance.

Table 1: Experimental results of semantic classification

Scene genre	Precision	Recall	F1
Conversation			
Emotional	0.937	0.914	0.925
Common	0.943	0.928	0.935
Action			
Gunfight	0.923	0.895	0.909
Beating	0.934	0.906	0.920
Chasing	0.893	0.932	0.912
Suspense	0.916	0.943	0.929

Table 2: Comparison of conceptually semantic classification

	Proposed			Yoshitaka *et al.* (1997)		
Scene	Pre.	Re.	F1	Pre.	Re.	F1
Conversation	0.940	0.921	0.930	0.859	0.867	0.863
Action	0.917	0.911	0.914	0.812	0.836	0.824
Suspense	0.916	0.943	0.929	0.873	0.847	0.860

Pre. and Re. denote precision and recall, respectively

CONCLUSIONS

In this study, we propose a novel scheme to classify video scenes into semantically meaningful categories based on the grammar of film to support high-level video retrieval in movie repository. Unlikely previous state-of-the-art literatures with scene definition which is not suitable for human descriptive language and features which can not fully describe the representative characteristics of scene content, the proposed scheme analyzes the characteristic structure information of semantically meaningful scene content using representative low and mid-level shot features which are robust and easily computable. We demonstrate the performance of the proposed approach by experimenting on Hollywood movies of various genres and the experimental results show the effectiveness of this scheme.

Future work includes the implementation of a more mature system. First, more types of conceptually meaningful scene content will be defined, e.g., argument, boxing. Second, more clips of various genres of movies will be conducted to further evaluate the performance of the proposed scheme, such as Documentary, Animation and so on. Finally, more representative features will be exploited to depict scene content. Context understanding through speech recognition is one of the possible schemes to be incorporated with the proposed system to further improve the accuracy of categorization.

ACKNOWLEDGMENTS

This study is supported by the Research Program of Nanjing University of Posts and Telecommunications under No. NY209018, NY209020 and NY208048. It is also supported by the Research Program under No. 08KJB510015 and YJCB2008039WL.

REFERENCES

Adams, B., C. Dorai and S. Venkatesh, 2000. Novel approach to determining tempo and dramatic story sections in motion pictures. Proceedings of the International Conference on Image Processing, Sept. 10-13, Vancouver, BC, Canada, pp: 283-286.

Bordwell, D. and K. Thompson, 1997. Film Art: An Introduction. 5th Edn., McGraw-Hill Publishing Company Inc., New York.

Huang, J.C., Z. Liu and Y. Wang, 2005. Joint scene classification and segmentation based on hidden Markov model. IEEE Trans. Multimedia, 7: 538-550.

Li, S., L. Zhu, Z.Q. Zhang, A. Blake, H.J. Zhang and H. Shum, 2002. Statistic learning of multi-view face detection. Proc. Eur. Conf. Comput. Vision, 2353: 67-81.

Li, Y., S.S. Narayanan and C.C. Kuo, 2003. Movie Content Analysis Indexing and Skimming. Chap. 5. Kluwer Academic Publishers, Video Mining, New York.

Pfeiffer S., R. Lienhart and W. Effelsberg, 2001. Scene determination based on video and audio features. Multimedia Tools Appli., 15: 59-81.

Sundaram, H. and S.F. Chang, 2000. Video scene segmentation using video and audio features. Proceedings of the IEEE International Conference on Multimedia and Expo, (ICME'00), Istanbul, Turkey, pp: 1145-1148.

Tavanapong, W. and J.Y. Zhou, 2004. Shot clustering techniques for story browsing. IEEE Trans. Multimedia, 4: 517-526.

Xie, L.X., P. Xu and S.F. Chang, 2004. Structure analysis of soccer video with domain knowledge and hidden Markov models. J. Pattern Recognition Lett., 25: 767-775.

Yang, Y. and X. Liu, 1999. A re-examination of text categorization method. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, New York, USA., pp: 42-49.

Yoshitaka, A., T. Ishii, M. Hirakawa and T. Ichikawa, 1997. Content-based retrieval of video data by the grammar of film. Proceedings of the 1997 IEEE Symposium on Visual Languages, Sep. 23-26, Faculty of Engineering, Hiroshima University, pp: 310-317.

Zhao, Y.J., T. Wang, P. Wang, W. Hu, Y.Z. Du, Y.M. Zhang and G.Y. Xu, 2007. Scene segmentation and categorization using NCuts. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 17-22, Minneapolis, MN., pp: 1-7.

Zhu, S.H. and Y.C. Liu, 2009a. Two-dimension entropy model for video shot partition. J. Sci. China Series F-Inform. Sci., 52: 183-194.

Zhu, S.H. and Y.C. Liu, 2009b. Automatic scene detection for advanced story retrieval. J. Expert Syst. Appli., 36: 5976-5986.