Film Narrative Exploration Through the Analysis of Aesthetic Elements

Chia-Wei Wang[1], Wen-Huang Cheng[2], Jun-Cheng Chen[1], Shu-Sian Yang[1], and Ja-Ling Wu[1,2]

[1] Department of Computer Science and Information Engineering
[2] Graduate Institute of Networking and Multimedia
National Taiwan University, Taipei, 10617, Taiwan, R.O.C.
{nacci,wisley,pullpull,pigyoung,wjl}@cmlab.csie.ntu.edu.tw

Abstract. In this paper, we propose a novel method for performing high-level narrative structure extraction of films. Our objective is to utilize the knowledge of film production for analyzing and extracting the structure of films. This is achieved by combining visual and aural cues on the basis of cinematic principles. An aesthetic model is developed to integrate visual and aural cues (aesthetic fields) to evaluate the aesthetic intensity curve which is associated with the film's narrative structure. Finally, we conduct experiments on different genres of films. Experimental results demonstrate the effectiveness of our approach.

1 Introduction

Film is one central part of the entertainment industry. Every year about 4,500 movies are released around the world, spanning over approximately 9,000 hours of digital movie contents, and the field is continuing to expand[1,2,6]. Since a film usually spans a long period of time and lacks organized metadata, extracting its content structures to facilitate user's access is a fundamental task in video analysis [7]. For film data, it is able to obtain the structures by analyzing the specific features called expressive elements (or aesthetic elements) that embedded in the film. Directors exploit the expressive elements to convey meanings, values and feelings during the production. Explicitly, directors create and manipulate expressive elements related to some aspects of visual or aural appeal to have perceptual or cognitive impact on the audience. Therefore, in this work, we utilize the knowledge of film production to analyze and extract the film structure.

1.1 Film Narrative Exploration

Narrative structure is the foundation upon which all stories are built to develop humans' cinematic literacy [1]. A classical narrative structure contains three basic parts called the beginning (exposition), the middle (conflict), and the end (resolution). The story intensity changes during different stages of the story. The term story intensity refers to the degree of tension that an audience feel about

T.-J. Cham et al. (Eds.): MMM 2007, LNCS 4351, Part I, pp. 606–615, 2007.

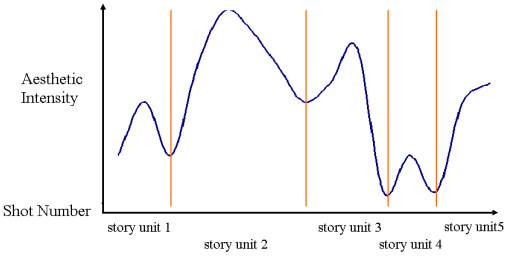

Fig. 1. A plot of story intensity curve and associated story boundaries

the story [4]. Generally, the story intensity is low at the beginning and then increases during the middle until reaches a climax. After the climax, the intensity diminishes at the end. In Figure 1, the three-part structure can be observed for each story unit. Later, this fact also helps to determine the corresponding story boundaries.

In film production, talented directors will purposely organize all movie shots to create a corresponding moods in a way that audiences will experience the same emotion enacted on the screen [8]. In addition, one of the director's major task is to emphasize a certain context or content, as for better expression of the situation of a story, in a manner such that audiences can naturally follow his way of story-telling. The storytelling now reflected through framing, light, color, objects, sounds, movement, and shot editing in a film [8]. A director applies the principle of media aesthetics to these basic aesthetic components to structure the visual and aural perception of a film [3]. For example, a director may use high energy colors to attract viewer's eyes and indicate the climaxes and emotional points of a story, etc. Therefore, directors can construct the aesthetic intensity structure that well corresponds to the story intensity structure [4].

Accordingly, it is able to detect and reconstruct such high-level mappings by extracting low-level computable features according to the principles of media aesthetics [3,9]. Zettl *et al.* [9] defined the media aesthetics as the study of how to apply the expressive elements to manipulate people's perception and helps media producers to translate significant ideas into significant messages efficiently, effectively, and predictably. Further, *Computational Media Aesthetics* proposed by Dorai *et al.* [3] provides a practical guidance for interpreting and evaluating expressive elements of films in an algorithmic way.

The rest of this paper is organized as follows. Section 2 illustrates the architecture of the proposed approach. In Section 3, we explain extractions of the adopted expressive elements and the associated aesthetic fields (light, color, movement, rhythm, and sound). The aesthetic model is presented in Section 4. Section 5 gives the experimental results and presents some applications. Finally, Section 6 concludes this paper and describes the directions of our future work.

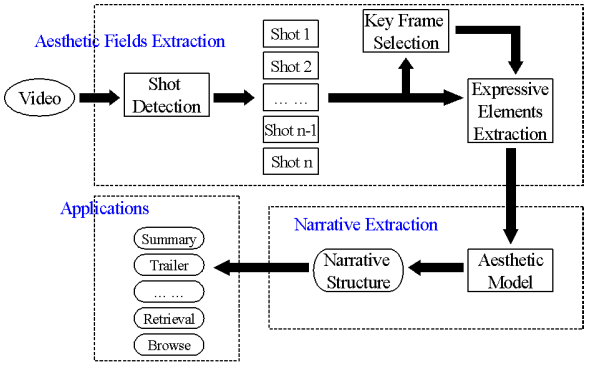

Fig. 2. The flowchart of the proposed algorithm

2 System Framework

We aim to extract the narrative structure of a film through computing the expressive elements used by film directors. Figure. 2 shows the flowchart of the proposed approach. The approach is composed of two stages: the aesthetic fields extraction and the narrative structure extraction through aesthetic modeling. In the aesthetic fields extraction stage, the first step is to explicitly detect the shot boundary between two consecutive frames. We compute the aesthetic fields associated with the expressive elements according to the principle of media aesthetics (see Section 3) on a keyframe basis. Next, in the stage of narrative structure extraction, we analyze the aesthetic fields extracted above. An aesthetic model is proposed (see Section 4) to evaluate the contribution of each field (denoted by the so-called aesthetic intensity) and obtain the narrative structure to realize some high-level video applications (see Section 5).

3 Aesthetic Field Extraction

According to the literatures [3,4,9], we identify and isolate five fundamental aesthetic fields (light, color, movement, sound, rhythm) that are computable and extractable for evaluating the aesthetic energy (intensity strength). The proposed framework is illustrated in Figure. 3. First, we compute the expressive elements, like color temperature and motion activity, directly from the keyframes of shots. Since the expressive elements themselves (such as shot length or motion activity) tell us nothing or little about the meaning expressed by directors, we further construct aesthetic fields by combining the extracted expressive elements. In this way, the so-obtained aesthetic fields are able to faithfully represent the semantic and perceptual importance of film events. Finally, we evaluate and combine the contributions of each aesthetic field and construct the aesthetic intensity structure through applying the aesthetic model to each shot. The adopted aesthetic fields are separately extracted for each keyframe and described as follows.

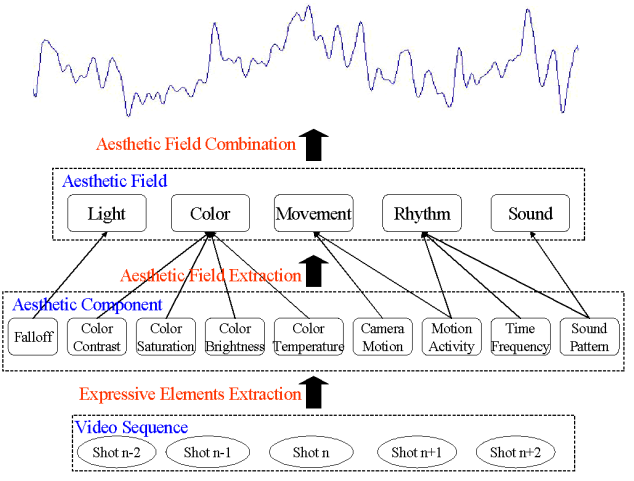

Fig. 3. The framework for extracting the aesthetic intensity curve

Light – It manipulates the perception of our environment and tells us how we would feel about a certain event. To structure the aesthetic field, light, many lighting instruments are generally used for the control of shadows than for just illuminating a scene. The brightness contrast between the light and the shadow sides of an object is referred as light falloff. To compute light falloff, we first coarsely classify the foreground and the background. Since the focused objects have more details within the object than the out-of-focus background, we adopt Wang's algorithm [12] to detect the focused objects in a frame using multi-resolution wavelet frequency method. After the classification of foreground and background, we use the idea of Mulhem *et al.* [11] to calculate the light falloff value. We calculate the luminance contrast along the boundary and linearly quantize the contrast values. Since the falloff edge often has the highest contrast, we use the average of the highest 10% contrast values along the edge as the light falloff value of the frame.

Color – It makes the audience feel in a specific way the content authors would like to communicate. We can translate colors into energies (or dynamics). The energy of a color presents the aesthetic impact on the audience, some colors seem to have high-energy and excite the audience while others seem to have low-energy and calm the audience down. Generally, it is common to use colors harmonically (high-energy color event matched by high-energy colors). The elements that influence the color energy are shown in [9].

Movement – It affects the viewers emotional reactions and impressions. When a human is watching a film, his emotional reactions and impression are often affected by the movement amount in the film. Generally, a larger movement will have greater visual intensity than a smaller one. We extract two kinds of movements (the object in front of the camera and the camera itself [4,9]) by

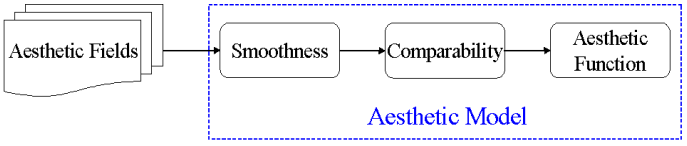

Fig. 4. The flowchart of aesthetic field modeling

using the motion activity descriptor defined in MPEG-7 [13]. The descriptor is compact and can be easily extracted in the compressed domain.

Sound – It helps to establish or supplement the visual effects of the screen event. Nonliteral sounds that refer to mainly the background and/or sound effects, can provide additional energy to a scene and quickly provide a desired mood. Since the semantic meaning and energy magnitude of literal sounds are not easy to measure, we focus on the energy of nonliteral sounds. We compute the loudness (volume) as the sound energy by the approximation of the root mean square value of the signal magnitude within a frame.

Rhythm – It is the perceived speed or felt time of an event [9]. For example, movement may produce a visual rhythm: when an actor slowly walk through a scene, the audience's felt time of this event is long and the rhythm is low; when the actor hurriedly run through the scene, the felt time is short and there is a high rhythm produced. Often the rhythm serves as a psychological guidance of audience. Generally, a faster (higher) rhythm is associated with excitement, and a slower (lower) rhythm suggests calm. Directors may control and present the rhythm by the techniques of montage (shot length), motion, and audio effects. We then adopt the formulation proposed in [14] to compute the rhythm elements.

4 Aesthetic Modeling

In this section, we explain in detail the process of evaluating the aesthetic intensity curve through integrating various aesthetic fields. Figure. 4 illustrates the procedure for modeling the aesthetic fields.

4.1 Smoothness

The aesthetic intensity of each aesthetic field is carefully smoothed via a smoothing window. The smoothing process is demanded for the following two reasons:

1) Memory is an important factor when considering the human perception. The emotional state of human is a function of a neighborhood of frames or shots and it does not change abruptly from one video frame to another.

2) Directors generally do not make the aesthetic intensity changing in a single or small number of shots. They often build a specific mood gradually from shot to shot.

In Hanjalic's original algorithm [10], a kaiser window is adopted to conduct the smoothing process. However, the memory is merely influenced by preceding

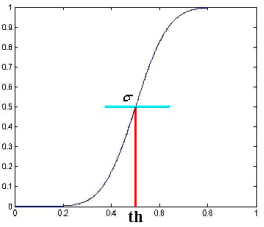

Fig. 5. Weighting function, the horizontal axis represents the value of the original curve while the vertical axis denotes the weighting value correspondingly

shots whereas the original kaiser window treats the preceding and posterior shots as equally important. Thus, we propose a modified kaiser window to reflect this property of human perception. To construct the modified kaiser window, two original kaiser windows are integrated together, both are of length 21, and the shape parameters are 3.5 and 31, respectively. We then combine the two kaiser windows into a new modified kaiser window which is then applied to conduct the smoothing process. Through the convolution with the modified kaiser window, we obtain the smoothed aesthetic intensity of each aesthetic field that takes account for the degree of memory retention of preceding frames and shots.

4.2 Comparability

This module ensures the aesthetic intensity of each aesthetic field is comparable and combinable. Each field is normalized by the shot with maximum magnitude in that field. Since the aesthetic intensities of all fields are scaled to the range between 0 and 1, they can be combined with each other on the same basis.

4.3 Aesthetic Function

As discussed previously, directors increase the energies or dynamics of aesthetic elements to emphasize the story conflicts or climax. According to the principle, we apply a filtering process to the intensity curve of each aesthetic field to provide highly distinguishable peaks at the segments of the curve corresponding to the story climax. The filtering process is performed through weighing the values of the aesthetic intensity curves. The weight function is defined as:

$$F(a(k), i = 1, ..., N) = \sum_{i=1}^{N} w_k a_k, \tag{1}$$

where

$$w_k = \frac{1}{2}(1 + erf(\frac{a_k - th}{\sigma})), \quad erf(x) = \frac{2}{\pi}\int_0^x e^{-t^2} dt. \tag{2}$$

As Figure. 5 depicts, the parameter **th** is the threshold of the original curve while σ denotes the spread factor determining the steepness of the middle curve

segment. The term a_k denote the aesthetic intensity curves that have been applied smoothing and normalizing processes as prescribed. The segments with intensity value larger than the threshold are strengthened while the other segments are lessened. Thus, we can distinctly identify the aesthetically emphasized segments and no possible climax segments are discarded.

5 Experiments and Applications

5.1 Experiments

In our experiments, five Hollywood movies are used, i.e., *Hitch (HT), Ghost (GH), Hero (HE), The Promise (TP),* and *Charlie's Angels II (CA)* to evaluate our performance. We compare the story units detected by our approach with those of both the ground truth taken from DVD chapter information and the human made results. Each of the videos was digitized at MPEG-1 format (352×240 pixels, 29.97fps). The data set is carefully selected to represent a variety of film genres such as action, drama, and romance. In addition, we examine the importance of each of the aesthetic fields.

Story Unit Segmentation. The chapter boundary detection is achieved based on the aesthetic intensity curve. As shown in Figure 1, a chapter boundary usually occurs at the point between the ending of the previous chapter unit and the opening of the next one. Since the chapter unit is usually with low aesthetic intensity at that point, we select the shot with the minimum intensity between two neighboring climax shots as the chapter boundary. We select those shots with the intensity value higher than a predefined threshold as the candidates for chapter unit climax since it can be found that the most impressive segments are often accompanied with a high aesthetic intensity values. Due to the fact that there is exactly one climax in a chapter unit and the shots near the climax are usually with higher values. For each pair of the candidate shots, if their distance is smaller than a threshold, ε_{high}, the shots with smaller intensity value are deleted from the candidate set.

Results. We compare the chapters detected by our approach with those of the ground truth (i.e., commercial DVD chapters, and manually labeled chapters). Table 2 and Table 3 show the statistics of our results as compared with the DVD chapter information and the manually labeled results, respectively. Note that a boundary would be identified as been correctly detected if it is within 35 seconds with a boundary in the ground truth. Since the chapter number in a commercial DVD is usually small to give viewers a rough idea about the video, it is reasonable that the overall recall is much higher than the precision. For real applications, the over-segmented chapters can be further grouped with further analysis. Overall, the experiment shows that our approach is successful in establishing the linkage between the computable aesthetic intensity and the abstract storyline of films.

Table 1. Comparisons with DVD chapters

Film	*HT*	*GH*	*TP*	*CA*	*HE*	Overall
Story units in ground truth	28	15	17	28	24	112
Story units detected	46	34	42	74	43	239
Correct detection	19	11	12	22	16	80
False negative	9	4	5	6	8	32
False positive	27	23	30	52	27	159
Recall	68%	73%	71%	79%	67%	71%
Precision	39%	29%	26%	28%	35%	33%

Table 2. Comparisons with Human Observers

Film	*HT*	*TP*	Overall
Story units in ground truth	36	30	66
Story units detected	44	42	86
Correct detection	26	23	49
False negative	10	7	17
False positive	18	19	37
Recall	72%	77%	74%
Precision	59%	55%	57%

Importance of Aesthetic Fields. We analyze the usefulness of each aesthetic field by removing one of the aesthetic fields at each time and re-evaluate the overall aesthetic intensity that is obtained from the reserved fields. For example, the weight parameter is 0.25 for each of the remaining four fields (note that the weight parameter is 0.2 for each of the five fields when no field is removed). From Table 3, it can be found that the overall performance drops while we remove any one of the aesthetic fields. These results show that it is essential to consider all of the aesthetic fields together.

Importance of Parameter. We also test other weighting schemes since each aesthetic field may not contribute equally to the human perception. Empirically, the weights of *rhythm*, *movement*, *sound*, *light*, and *color* are set to 0.2, 0.23, 0.11, 0.26, and 0.2, respectively. The results are shown in Table 4. It demonstrates that there is a notable gain in performance after tuning the weights. Besides, different film genres possess different art forms, and a certain weights may work the best for a particular film genre. For example, action films have more motion activity and faster rhythm than those of the other genres. The performance can be improved if taking this fact into account. We analyze each aesthetic field of *Charlie's Angel II* (an action movie) against the corresponding DVD chapter information. We found that the sound and the light fields do not work well and we decrease their weights. Empirically, the weights of *rhythm*, *movement*,

Table 3. Importance of different aesthetic fields

Feature (removed)	*rhythm*	*movement*	*sound*	*light*	*color*
Recall	63%	63%	62%	63%	49%
Precision	30%	30%	28%	31%	26%

Table 4. Performance gains from adjusting weights

Type	Linear	Tuned
Story units in ground truth	112	112
Story units detected	239	242
Correct detection	80	85
False negative	32	27
False positive	159	157
Recall	71%	76%
Precision	33%	35%

Table 5. Performances of different films for a given set of weights

Film	*CA*	*HT*	*GH*	*TP*	*HE*
Recall	+13.6%	-15.8%	-27.3%	-8.3%	-12.5%
Precision	+12.1%	-15.8%	-31.3%	-10.5%	-21.6%

sound, *light*, and *color* are set to 0.2, 0.22, 0.14, 0.14, and 0.3, respectively. The performances of each film under the given weights are listed in Table 5. There is a remarkable performance gain in *Charlie's Angel II* while the performances of the other films drop drastically. Therefore, automatic weights selection for different film genres is an important issue and will be the major direction of our future work.

5.2 Applications

As described in [5], identification and extraction of the high-level narrative structure associated with the expressive elements and the form of story in films opens the way for more sophisticated applications to meet the demands of the audience. For example:

1) It helps to automatically generate video indexes and makes it possible for query specification in semantic terms such as "Where is the most intense part of the movie?" or "How long is the first story unit last?", etc. Generally speaking, the higher the level of the structure is, the more efficient the search would be.

2) It locates the important boundaries of a film or a story segmentation to meet viewers' need to gain more control of what they see, e.g., DVD options are being made for users to randomly view a specific story unit of the movie.

3) It enables us to give the summaries of movies for efficiently browsing and previewing the movie.

6 Conclusion and Future Work

We proposed a method to perform high-level narrative structure extraction of films. We demonstrate that combining visual and aural cues with the aid of cinematic principles can provide significant performance for extracting the corresponding narrative structure. In the future, we are interested in concatenating small story units into longer and more meaningful ones for further applications.

References

1. N. Abrams, I. Bell, and J. Udris, *Studying Film.* London: Hodder Headline Group and NY: Oxford University Press, 2001.
2. J. Monaco, *How to Read a Film, 3ed.* NY: Oxford University Press, 2000.
3. C. Dorai and S. Venkatesh, *Media Computing: Computational Media Aesthetics.* Boston/Dordrecht/London: Kluwer Academic Publisher, 2002.
4. B. Block, *The Visual Story: Seeing the Structure of Film, TV, and New Media.* Boston: Focal Press, 2001.
5. B. Adams, C. Dorai, S. Venkatesh, and H. H. Bui, "Indexing narrative structure and semantics in motion pictures with a probabilistic framework," *IEEE International Conference on Multimedia and Expo* (ICME'03), vol. 2, pp. II 453-456, July 2003.
6. Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Transactions on Multimedia,* vol. 7, no. 6, pp. 1097-1105, Dec 2005.
7. R. Yong, S. H. Thomas, and S. Mehrotra, "Constructing table-of-content for videos," *Multimedia Systems,* vol. 7, pp. 359-368, Sept 1998.
8. R. W. Picard, *Affective Computing.* MA: The MIT Press, 1997
9. H. Zettl, *Sight, Sound, Motion: Applied Media Aesthetics.* SF: Wadsworth, 1973.
10. A. Hanjalic, and L. Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia,* vol. 7, no. 1, pp. 143-154, Feb 2005.
11. P. Mulhem, M. S. Kankanhalli, Y. Ji, and H. Hassan, "Pivot Vector Space Approach for Audio-Video Mixing," *IEEE Multimedia,* vol. 10, pp.28-40, April-June 2003.
12. J. Z. Wang, J. Z, R. M. Gray, and G. Wiederhold, "Unsupervised multiresolution segmentation for images with low depth of field," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 1, pp. 85-90, Jan 2001.
13. S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptors," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 11, no. 6, pp. 720-724, June 2001.
14. H. W. Chen, J. H. Kuo, W. T. Chu, and J. L. Wu, "Action movies segmentation and summarization based on tempo analysis" *Proc. ACM SIGMM International Workshop on Multimedia Information Retrieval* (MIR'04), pp. 251-258, Oct 2004.