Automatic Highlights Extraction for Drama Video Using Music Emotion and Human Face Features

Keng-Sheng Lin[1], Ann Lee[2], Yi-Hsuan Yang[1], Cheng-Te Lee[3], and Homer H. Chen[1]

[1]*Graduate Institute of Communication Engineering* [2]*Department of Electrical Engineering* [3]*Graduate Institute of Computer Science and Information Engineering*

National Taiwan University
Taipei, Taiwan

[pridek0912, an918tw, affige, aderleee]@gmail.com, homer@cc.ee.ntu.edu.tw

Abstract—**The rich emotion part of a drama video is often the center of attraction to the viewer. Emotion-based highlights extraction is useful for applications such as drama video retrieval and automatic trailer generation. In this paper, we propose a system that uses music emotion and human face as features for automatic extraction of the emotion highlights of a drama video. These high-level audiovisual features are used because music invokes emotion response from the viewer and characters express emotion on their faces. To avoid the interference of speech signal and environmental noise, a novel two-stage music emotion recognition scheme is developed. We first detect the presence of incidental music in a drama video using an audio fingerprint technique, and then perform emotion recognition on the noise-free music available from the album of the incidental music. This simple but effective approach greatly improves the accuracy of music emotion recognition. Besides the conventional subjective evaluation, we propose a new metric for quantitative performance evaluation of highlights extraction. Evaluation results are provided to illustrate the performance of the system.**

I. Introduction

With the proliferation of multimedia data, converting long videos into short ones for more effective video retrieval or browsing is needed [1]. There are two different approaches: summarization and highlight. The former aims to provide a condensed and succinct storyline representation of a video, whereas the latter aims to extract the affective content from the video. Recent studies show that humans tend to remember affective content easily [2]. Thus, it is the primary interest of this paper to investigate highlights extraction, especially for drama video—one of the most popular genres of TV programs.

Here, affective content refers to audio or visual segments that invoke strong emotion responses such as laughing or fear from the viewer [3], and drama video refers to television series presented in the form of a number of episodes and containing audio and visual data. Previous studies on affective content analysis mainly focused on movie. Although a director can narrate a story through drama video or movie, the techniques used to present affect are somewhat different because of factors such as frame resolution, length of video, audience etc. A drama video is also different from sitcom [4] in that it normally has shots of the physical place, for example an airport, where a dramatic incident takes place. Therefore, there are specific characteristics of drama video that can be exploited for better highlights extraction.

Hanjalic and Xu [5] are the first who described the affect of a video by an affect curve on a two-dimensional arousal-valence plane developed by Thayer [6]. Low-level audiovisual features such as motion vector, shot duration, and sound energy were mapped to an arousal value, and average pitch of the audio data to a valence value. However, due to the semantic gap between the audiovisual features and the human emotion perception, the accuracy of highlights extraction purely based on low-level features is limited. Other low-level features such as color [7], emotion-related mid-level features such as laughing and screaming [8], and emotion-related key words such as love and hate in the subtitle [8] can be applied to improve the performance of highlights extraction.

It has been observed in psychology and filmology studies that human face conveys crucial information (for example, direction of eye gaze) for social interaction [9] and that music can invoke emotion response from the listener [10], [11]. In fact, directors of drama video often bring the viewer's emotion to a climax through the use of incidental music, and that the presence of actors and actresses in a sequence of consecutive frames is often indicative of a pinnacle moment of the drama (the longer the sequence, the stronger the highlight). Therefore, we use face and music emotion to improve the accuracy of highlights extraction and incorporate these two high-level features in our system.

We also investigate the relation between highlights and two low-level visual features, shot duration and motion magnitude, which are exploited by directors to generate emotion effects. Specifically, short shot duration and high motion magnitude are used to highlight an action scene, whereas long shot duration and low motion magnitude are used to highlight a romantic scene [10].

In short, the contribution of this paper is three-fold. First, we present a system that integrates information from music emotion, human face, shot duration, and motion magnitude for highlights extraction of drama video (Sections II–IV). Second, we devise a novel two-stage music emotion recognition scheme and a novel adaptive audio fingerprint technique to improve the accuracy of emotion recognition for incidental music (Section II). Third, we propose a novel distance metric that can be utilized to evaluate the performance of highlights extraction in a quantitative fashion (Section V). To our best

978-1-4577-1434-4/11/$26.00 ©2011 IEEE

knowledge, few quantitative evaluations of highlights extraction if any have been reported in the literature.

II. Music Detection and Music Emotion Recognition

The input drama video consists of audio and visual signals, which are processed separately. For the audio signal, the system detects the presence of incidental music and employs the well-known MIRtoolbox [12] for music emotion recognition of the incidental music. As it is usually the case, we assume that the album that contains all the incidental music used in a drama video, a.k.a. the original soundtrack, is available when the drama video is released. The music emotion recognition is performed on the music provided in the album, not the input audio signal. Emotion recognition is performed on the clean data because the input audio signal may be corrupted with speech signal and environmental noise, which usually degrade the accuracy of emotion recognition. Furthermore, given an album, we use audio fingerprint [13] to detect the presence of each incidental music and the specific portion an incidental music is played. Another advantage of the audio fingerprint approach is that it is free of the time-consuming and labor-intensive labeling of the training data that is typically required for conventional machine learning approach [14].

A. *Adaptive Music Detection by Audio Fingerprint*

As a content-based signature, audio fingerprint has been used to characterize or identify an audio sample [15]. For example, the audio searching engine developed by Wang [13], [16] applies a short-time Fourier transform to an audio segment and chooses local peaks as landmarks. The differences of time and frequency values (expressed by a hash table) between landmarks of neighboring time windows constitute the fingerprint. The more matching between two audio segments in the hash values, the more likely the two segments are originated from the same song. Since the landmarks have high energy relatively, the audio fingerprint is robust to noise.

The similarity of two audio segments is indicated by the number of matched hash values. In our case, one of the two audio segments represents an input audio segment, and the other represents an audio segment of an incidental music in the album. Then a binary decision is made according to the similarity score to determine whether the input audio segment is music or not. The threshold used for the binary decision is of fundamental importance, because it controls the accuracy of music detection. However, since the audio signal of drama video is a blend of speech and music signals, the similarity between a pure music segment and a blended audio segment is not as high as that of two pure music segments. Such pair of audio segments can be erroneously classified as dissimilar if a constant threshold is used. For highlights extraction, we want to detect all music segments including those that are blended with speech.

We propose an adaptive technique that automatically determines the threshold through the use of a low short-time energy ratio ρ of the input audio segment. The low short-time energy ratio ρ is obtained by counting the number of frames with short-time energy e smaller than one half of the average short-time energy $\bar{e}$ and dividing the resulting number by N, the total number of frames in a time window. Specifically,

$$\rho = \frac{\sum_{n=0}^{N-1}[sgn(0.5\bar{e}-e_n)+1]}{2N}, \quad (1)$$

where $sgn(\cdot)$ is the sign operator that yields 1 for positive input and –1 for negative input, and e_n is the short-time energy of the nth frame. The short-time energy is computed on a frame basis. If an audio signal has a high ρ value, it has more silence frames. In general, speech signals have more silence frames than music signals. Therefore, we can discriminate speech from music based on the ρ value.

Table I.
Adaptive threshold associated with the low short-time energy ratio

ρ	Threshold
0.0–0.2	3
0.2–0.4	4
0.4–0.6	5
0.6–0.8	6
0.8–1.0	7

Then the adaptive threshold is determined empirically by observing the relationship between the value of ρ and the number of matched hash values. The ρ value and its corresponding threshold are shown in Table I, where the range of ρ is partitioned evenly into 5 bands, and a different threshold value for each band is assigned. As we can see, the threshold increases with the ρ value. This is desirable because when the input audio is more likely to be a speech signal as judged by its ρ value, we raise the threshold to avoid possible false matches and thereby achieve better accuracy in music detection.

B. *Music Emotion Recognition*

Instead of representing emotions as discrete labels such as anger and happiness, we approach it from a dimensional perspective and define emotions as points in a three-dimensional space [17], the three axes of which are arousal (exciting or calming), valence (positive or negative), and dominance (a sense of control or freedom to act). This dimensional approach avoids the ambiguity and granularity issues inherent to discrete labels [18]. In addition, it allows one to intuitively treat emotion points that are far away from the origin in the 3-D emotion space as emotions of high intensity [19]. Based on this framework, we apply machine learning models to predict the emotion values of each short-time music segment with respect to the three dimensions and consider the Euclidean distance between the predicted emotion values and the origin as the highlight score of the music segment.

Specifically, we apply the MIRtoolbox [12] to predict the emotion value of a music segment. The training dataset of the emotion value prediction module is composed of 110 audio sequences of film music (each sequence is on the average

Fig. 1. Two highlight frames of the drama *Flower Shop without Rose*

annotated by 116 human subjects). A number of music features (timbre, harmony, rhythm, etc.) are extracted, and a regression algorithm called multivariate regression [20] is employed to learn the relationship between music features and emotion values. One regression model is trained for each emotion dimension. As the characteristics of film music are close to the music used in drama videos, we believe the MIRtoolbox is applicable to our system.

III. Visual Features Extraction

It has been shown that the presence of human face in video often catches the viewer's attention and invokes emotion [9], [21]. Shot duration and motion magnitude also impact the affective response of a viewer [5]. This section describes how our system uses human face, shot duration, and motion magnitude as features for highlights extraction of drama video.

A. Human Face

The use of face as a feature for highlights extraction in our system is based on two main observations. First, the highlight scenes of a drama video often show the interactions between characters. Therefore, human face represents an important candidate feature for highlights extraction. Second, in the context of video highlights, the size of face does not correlate well with its importance. For example, the two frames shown in Fig. 1 represent the highlights of a TV drama, but the size of human face is obviously quite different. Therefore, we direct our attention to the number of human faces appear in each frame. Typically, when more faces appear in a frame, there are more interactions between the characters, and this attracts more viewers' attention. However, faces that are too small (for example, less than 5% of the frame size) to invoke affective response are not considered. In our system, we adopt the algorithm described in [22] for face detection.

The highlight score of human face is obtained as follows. Each frame is initially assigned a score equal to the number of faces detected for the frame. An example sequence of initial scores is shown in Fig. 2. Then the initial score of each frame is propagated to its two neighboring frames. Finally, the initial score of each frame is summed with the initial scores of the two neighboring frames to form the final score (called smoothed score because the summation is in essence a temporal smoothing operation).

B. Shot Duration

Shot duration is used as a feature in our system for highlights extraction. It has been found that short-duration shots can invoke higher arousal than other shots [5]. This accounts for the fact that short-duration shots are often part of the highlight scenes of a drama video. On the other hand, directors often use long-duration shots to draw the viewer's attention to specific scenes such as romance and slow motion [10]. This is why one can also find long-duration shots in a highlight scene as well. On the basis of these observations, we propose to detect both short- and long-duration shots in the highlights extraction process.

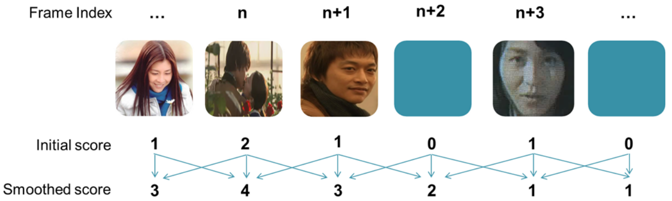

Fig. 2. An illustration of the calculation of highlight score of human faces for the drama video *Flower Shop without Rose*

In our system, the strength of a shot as a candidate of the highlights is modeled by an exponential function. Let the highlight strength of a frame in the *k*th shot be denoted by s_k and the shot duration by $n_k - p_k$, we have

$$s_k = e^{1-(n_k-p_k)}, \qquad (2)$$

where p_k is the index of the first frame and n_k the last frame of the *k*th shot. As we can see, the shorter the shot duration, the higher the s_k. To have a high score for both short- and long-duration shots, we measure the highlight score of a shot duration by its distance from the average highlight strength. More precisely, the highlight score $\hat{s}_k$ of any frame in the *k*th shot is obtained by

$$\hat{s}_k = |s_k - \bar{s}|, \qquad (3)$$

where $\bar{s}$ is the average of s_k's of an episode.

C. Motion Magnitude

A motion vector contains both magnitude and orientation information. Unlike motion orientation, motion magnitude is a good indication of highlight. Similar to the case for shot duration, a highlight scene of drama video may not strictly contain fast-motion frames; slow-motion frames can be part of the highlights as well. For example, a scene showing tears slowly trickling down an actress's face can invoke strong emotion responses from the viewers [10]. Therefore, we propose to consider both fast- and slow-motion frames in the highlights extraction process.

The detection of fast- and slow-motion frames consists of two steps. First, we compute the normalized average motion magnitude a_k of all blocks within the *k*th frame,

$$a_k = \frac{\sum_{i=1}^{I}|\vec{v}_k(i)|}{I|\vec{V}_k|}, \qquad (4)$$

where $\vec{v}_k(i)$ is the motion vector of the *i*th block, I is the total number of motion vectors, and $\vec{V}_k$ is the largest motion vector of the frame. As we can see, the average motion is normalized with respect to the largest motion vector.

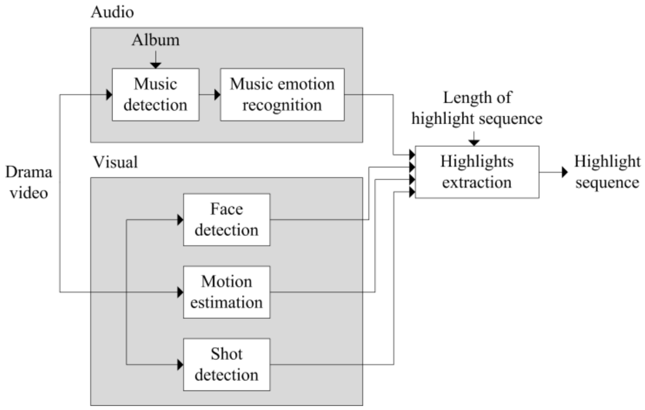

Fig. 3. The proposed highlights extraction system for drama video

Second, to signify both fast- and slow-motion frames and suppress average-motion frames, we measure the highlight score $\hat{a}_k$ of the kth frame of a drama video by

$$\hat{a}_k = |a_k - \bar{a}|, \tag{5}$$

where $\bar{a}$ is the average of a_k's of an episode. This method turns out to be effective yet simple for sifting video frames with either high or slow motion.

IV. Highlights Extraction Procedure

Fig. 3 shows the block diagram of the proposed system. Given a drama video, the system processes the input audio and video signals separately. It first extracts the audio-visual features and computes the highlight score of each feature using the methods described in Sections II and III. Then it linearly combines the highlight scores to obtain the overall highlight score, denoted by *H*, for each second of the drama video. Finally, it extracts those video segments with highlight score higher than a threshold. The threshold is automatically determined by the system according to the desired length of the highlight sequence.

The overall highlight score is computed by a weighted sum of the four highlight scores

$$H = f_M H_M + f_F H_F + f_S H_S + f_A H_A, \tag{6}$$

where H_M, H_F, H_S, and H_A, respectively, are the scores of music emotion, human face, shot duration, and motion magnitude and f_M, f_F, f_S, and f_A are the corresponding weighting factors. In our system, each of these four scores is normalized to the range [0, 1] and sum-to-one. In our experiments, the four weights are simply set to [0.3, 0.3, 0.2, 0.2], respectively, to place more emphasis on high-level features over low-level ones. Our evaluation also shows that high-level features are indeed more useful for highlights extraction (c.f. Section VI.*A*).

V. Distance Metric for Highlights

In our experiment, the consistency between the highlight sequence generated by the proposed system and the subjective results provided by the subjects for the same drama video are evaluated. A high consistency means the system performs well. Each subject expressed in writing their view of the highlights of a drama video after watching the entire drama video. More precisely, we ask the subjects to describe, in words, the highlights at the story unit level (for example, "actor knees down and propose to actress") and give a score to each highlight. Similarly, the score of each video segment that is extracted by the proposed system is the summation of the *H* within the segment.

Story unit [23] in hierarchical video data representation is a series of shots that communicate a unified action with a common locale and time. Because it is easier for the subjects to perceive and remember a video at the level of story unit [24], especially when the drama video is long, we ask them to describe the highlights of each story unit than to label the video segments. Because the result given by the subjects is in the form of textual description, the result is manually interpreted and matched with video data. Continue the previous example. If there is indeed a knee-down scene in a certain video segment of the highlight sequence, then the textual description is considered a match with the video data. After finding the matched video data, we compute the difference of the scores given by the proposed system and the subject.

To measure the consistency, we compute the distance *D* between the highlight sequence and the textual description provided by the subjects according to the following equation,

$$D = \sum_{\mathbb{s} \in \mathbb{S}_A \cup \mathbb{S}_B} |w_A(\mathbb{s}) - w_B(\mathbb{s})|, \tag{7}$$

where *A* is the highlight sequence, *B* is the textual descriptions of the highlights, $\mathbb{S}_A$ is the set of story units of *A*, $\mathbb{s}$ is the story unit belongs to $\mathbb{S}_A \cup \mathbb{S}_B$, and $w_A(\mathbb{s})$ returns the score of the story unit $\mathbb{s}$ of *A*. Note that we define $w_A(\mathbb{s})$ to be zero if $\mathbb{s}$ is not belongs to *A*. The total score of the highlight sequence (the textual description) is normalized to one. The value of *D* is between 0 and 2; it is equal to 0 when *A* and *B* are exactly identical.

VI. Experimental Results

We conduct three tests. In the first test, we investigate the effectiveness of the four features that are used in the proposed system by using the distance metric described in Section V. In the second test, we evaluate the effectiveness of the proposed system by the distance metric described in Section V. In the third test, subjects are asked to evaluate how strong their emotion is invoked by the highlights generated by the system. In this way, we can gain further insights into the overall impressions of video highlights. In the three tests, the length of highlight sequence for each drama video is close to 33 minutes (1980±10 seconds).

Although many systems for video highlights extraction or summarization have been developed, most of them focus on sports video or news video [25]. Since the contents as well as design considerations are different, it makes little sense to compare such systems with ours. Instead, in the second and third tests, the proposed system is compared with a baseline

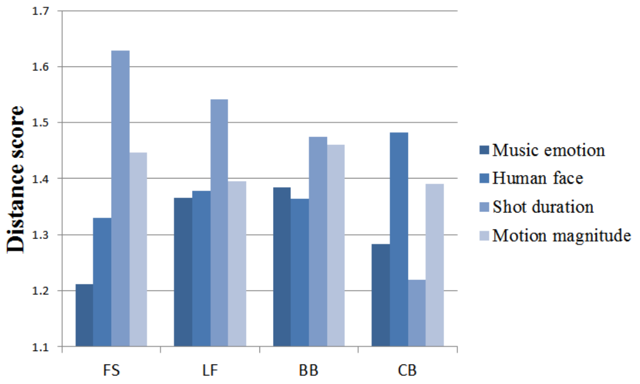

Fig. 4. Comparison of the effectiveness of the four features on four drama videos

system that generates highlight sequence by uniformly sampling one-minute video segments in a video. Such a baseline system has been widely used for video browsing and is easy to be implemented [26].

The test data set contains four drama videos: *Flower Shop without Rose* (denoted by FS), *Last Friends* (LF), *Buzzer Beat* (BB), and *Code Blue* (CB). Different types of stories are depicted in the four videos; for example, FS is a romantic drama and CB is a medical drama. Textual descriptions for the highlights of the four drama videos are collected from 13, 13, 16, and 12 subjects, respectively.

A. Quantitative Evaluation of Individual Feature

We generate four highlight sequences by using each of the four features alone; that is, the weighting factor of the feature used is one and the others are zero. The distance D between the highlight sequence and the highlights provided by each subject is computed.

Fig. 4 shows the average of distance D on the four drama videos. We can see that the high-level features, music emotion especially, are more effective for highlights extraction than low-level features. However, in CB (a medical drama), human face is not as effective as it in the other drama videos. The reason could be that the highlight scenes of medical dramas are often the surgical scenes where actors playing doctors always wear masks. This degrades the accuracy of face detection and thus influences the effectiveness of human face.

B. Quantitative Evaluation of the Integrated System

The distance D between the highlight sequence and the highlights provided by each subject is computed. Note that the video segments extracted by the baseline system get the same score because they are simply a uniform sampling of the video. The oracle system represents a lower bound of distance on using the set of the highlights provided by the subjects.

Fig. 5 shows the means and standard deviations of distances between the highlights outputted by three different approaches and the highlights provided by the subjects for the four videos. There are two observations to be made. First, the distance of the proposed system that uses the four features are less than that uses only one feature; in other words, the four features are complementary to detect highlights from drama video. Second, the distances of the proposed system are less than those of the baseline system, and the improvement is significant under the one-tailed t-test (p-value<0.05) [27]; that is, the highlight sequences generated by the proposed system are more consistent with the subjective results provided by the subjects.

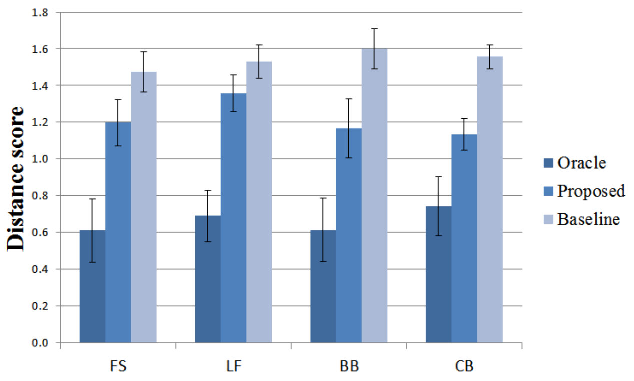

Fig. 5. Comparison of the performance of the oracle system, the proposed system, and the baseline system on four drama videos

TABLE II.
(MEAN, STANDARD DEVIATION) VALUES OF SUBJECTIVE TEST

Drama	Subjects of the first group		Subjects of the second group	
	The proposed system	**The baseline system**	**The proposed system**	**The baseline system**
FS	(7.3, 1.05)	(6.2, 2.04)	(7.5, 1.06)	(5.9, 1.28)
LF	(7.3, 1.22)	(5.4, 1.75)	(7.6, 1.14)	(5.8, 1.20)
BB	(7.1, 0.92)	(5.1, 1.39)	(6.8, 1.21)	(5.8, 1.42)
CB	(7.7, 0.99)	(6.2, 1.31)	(7.7, 0.73)	(5.9, 1.17)

C. Qualitative Evaluation of the Integrated System

We ask the subjects watch the highlight sequences that are used in the second test and then give their subjective scores to evaluate how strong their emotion is invoked by the highlight sequences. The score ranges between 1 and 10. The higher the score, the stronger the subject's emotion is invoked. Because the watching experience might influence the results of evaluation, the subjects are divided into two equal-sized groups. The subjects of the first group have seen the entire drama before and the subjects of the second group have not. There are 30, 32, 26, and 28 subjects for the four drama videos. The results are listed in Table II.

Table II shows that both groups of subjects agree that the proposed system effectively extract highlights of drama video. The improvement over the baseline system is significant under the one-tailed t-test (p-value<0.05).

VII. CONCLUSION

Video highlights extraction often requires domain-specific knowledge. Despite that this subject has been intensively studied, little work has focused on the video highlights extraction for drama video. In this paper, we use two high-level features (music emotion and human face) and two low-level features (shot duration and motion magnitude) to extract highlights based on the findings of psychology and filmology studies. Using the noise-free incidental music album allows us to prevent the effect of noise on music emotion classification and thereby improve the performance of overall system. The evaluation results successfully illustrate that the high-level features, especially music emotion, are effective for video highlights extraction.

References

[1] A. Hanjalic, N. Sebe, and E. Chang, “Multimedia content analysis, management and retrieval: Trends and challenges,” *Proc. SPIE-IS&T Electronic Imaging*, vol. 6073, pp. 1–5, 2006.

[2] A. Lang, J. Newhagen, and B. Reeves, “Negative video as structure: Emotion, attention, capacity, and memory,” *J. Broadcast and Electron. Media*, vol.40, pp. 460–477, 1996.

[3] R. W. Picard, “*Affective Computing*,” MIT Press, Cambridge, MA, 2000.

[4] M. Lewisohn, "*Radio Times Guide to TV Comedy*," BBC Worldwide, 2003.

[5] A. Hanjalic and L. Q. Xu, “Affective video content representation and modeling,” *IEEE Trans. Multimedia*, vol. 7, no. 1, February, 2005.

[6] R. E. Thayer, “*The Biopsychology of Mood and Arousal*,” Oxford University Press, 1989.

[7] C. W. Ngo, Y. F. Ma, and H. J. Zhang, “Video summarization and scene detection by graph modeling,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, no. 2, February, 2005.

[8] M. Xu, S. Luo, J. S. Jin, and M. Park, “Affective content analysis by mid-Level representation in multiple modalities,” *Proc. Internet Multimedia Computing and Service*, 2009.

[9] R. Palermo and G. Rhodes, “Are you always on my mind? A review of how face perception and attention interact,” Neuropsychologia, vol. 45, no. l, pp. 75-92, 2007.

[10] L. D. Giannetti, “*Understanding Movies*,” 10th edition, Prentice Hall, 2004.

[11] M. Shevy, “The mood of rock music affects evaluation of video elements differing in valence and dominance,” *Psychomusicology: Music, Mind, and Brain,* vol. 19, no. 2, 2007.

[12] O. Lartillot, P. Toiviainen, and T. Eerola, “A Matlab toolbox for music information retrieval,” in C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.), *Data Analysis*, *Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2008.

[13] A. Wang, “An industrial-strength audio search algorithm,” *Proc. International Society for Music Information Retrieval*, 2003.

[14] A. Temko and C. Nadeu, “Classification of acoustic events using SVM-based clustering schemes,” *Pattern Recognition*, vol. 39, pp. 682–694, 2006.

[15] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, “A review of audio fingerprinting,” *J. VLSI Signal Processing*, vol. 41, no. 3, pp. 271–284, 2005.

[16] D. Ellis, “Robust landmark-based audio fingerprinting”, web resource, available: http://labrosa.ee.columbia.edu/matlab/fingerprint/

[17] A. Mehrabian, “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” *Current Psychology*, col. 14, no. 4, pp. 261–292, 1996.

[18] Y. H. Yang, Y. C. Lin, Y. F. Su and H. H. Chen, “A regression approach to music emotion recognition,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[19] H. L. Wang and L. F. Cheong, “Affective understanding in film,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 16, no. 6, June, 2006.

[20] T. Eerola, O. Lartillot, and P. Toiviainen, “Prediction of multidimensional emotional ratings in music from audio using multivariate regression models,” *Int. Conf. Multimedia Information Retrieval*, 2009.

[21] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," *ACM Int. Conf. Multimedia*, 2010.

[22] W. Kienzle, G. Bakir, M. Franz, and B. Scholkopf, “Face detection - efficient and rank deficient,” *Advances in Neural Information Processing Systems*, pp. 673–680, 2005.

[23] J. M. Boggs and D. W. Petrie, “*The Art of Watching Films*”, 5th edition, Mountain View, CA: Mayfield, 2000.

[24] Y. Rui, T. S. Huang, and S. Mehrotra, “Constructing table-of-content for videos,” *Multimedia System, Special Section on Video Libraries*, vol. 7, no. 5, pp. 359–368, 1999.

[25] A. G. Money and H. Agius, “Video summarization: A conceptual framework and survey of the state of the art,” *J. Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008.

[26] L. Herranz and J. M. Martinez, “A framework for scalable summarization of video,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 9, September, 2010.

[27] M. Hollander and D. A. Wolfe, “*Nonparametric Statistical Methods*,” John Wiley & Sons Inc., Hoboken, NJ, 1999.