Estimating Multiple Evoked Emotions from Videos

Wonhee Choe (wonheechoe@gmail.com)
Cognitive Science Program, Seoul National University,
Seoul 151-744, Republic of Korea
Digital Media & Communication (DMC) R&D Center, Samsung Electronics
Suwon 443-742, Republic of Korea

Hyo-Sun Chun (hschun@bi.snu.ac.kr)
Junhyug Noh (jhroh86@gmail.com)
School of Computer Science and Engineering, Seoul National University,
Seoul 151-744, Republic of Korea

Seong-Deok Lee (lsdlee@samsung.com)
Future IT, Samsung Advanced Institute of Technology (SAIT),
Yongin 449-712, Republic of Korea

Byoung-Tak Zhang (btzhang@bi.snu.ac.kr)
Computer Science and Engineering & Cognitive Science and Brain Science Programs,
Seoul National University, Seoul, 151-744, Republic of Korea

Abstract

Video-sharing websites have begun to provide easy access to user-generated video content. How do we find what we want to view among the huge video database? When people search for a video, they may want to know whether the video evokes a certain emotional sensation. The evoked emotion is one of the important factors we consider when we select a video. One of the key concepts of evoked emotions from videos: the evoked emotions are different for each scene and for each viewer. Considering these differences, we obtained human-evoked emotions from 33 videos. We used these emotions to estimate the multiple emotions evoked by each scene of the videos. Using a computational model of emotion estimation based on mid-level visual features, we found that, in individual videos, the same scene evoked multiple emotions. Our results show that a video evoked different emotions from different people. A computational model might deliver probabilistic multiple-evoked emotions from video analyses.

Keywords: evoked emotion; visual feature; video retrieval

Introduction

Video-sharing websites provide easy access to a wide variety of user-generated video content, including movie clips, television clips, music videos, and amateur content. We can search a huge database of videos for a video that we want to see via the use of keywords or a search by genre. Unfortunately, these search strategies are not sufficient. If we do not have any prior knowledge of a video, how do we find what we want to see? The mood of movie is one of the most important factors we consider when we select a movie. When people search for a movie or TV series, they may want to know whether the video has a mood that is similar to that of a video they have viewed before. Sometimes, they may want to change their mood by watching a video.

Until now, most of the research efforts have focused on a content- or genre-based analysis of videos despite many users' needs for emotion-sensitive video retrieval. Fortunately, some have studied the emotions evoked by images (Wang & He, 2008; Yanulevskaya et al., 2008; Li, Zhang & Tan, 2010). Wang and He (2008) focused on emotional semantic image retrieval (ESIR) instead of content-based image retrieval (CBIR) to reduce the "semantic gap." Others studied emotional picture categorization using the International Affective Picture System (IAPS) according to the 10-emotion model (Yanulevskaya et al., 2008; Li, Zhang & Tan, 2010). However, the application of these approaches to a video sequence may not be appropriate due to the lack of consideration of temporal variations. Various moods change sequentially within a given video. That is, within a given video, different emotions may be evoked across scenes.

Two different studies evaluated the emotions evoked by videos. Canini et al. (2009) evaluated the emotional identity of videos. The research used light source color, motion dynamics, and audio track energy as the temporal features of videos. It was a first attempt to evaluate the temporally changing emotional identity of movies. They used a 3 dimensional (3D) emotional identity space (warm/cold, dynamic/slow, and energetic/minimal) to show the trajectory of one video clip. Unfortunately, the emotional identity space was used to express movies' content changes rather than humans' various emotional changes. Bailenson et al. (2008) focused on a classification algorithm of emotions evoked by videos. Facial feature tracking was used and physiological responses were measured. The study tried to predict 2 emotions (amusement and sadness) and the intensity of each emotion. Unfortunately, this approach is not applicable to the study of video retrieval.

Winter and Kuiper (1997) reported that individual difference factors play an important role in the experience of emotion. In fact, individuals may respond differently depending on their current state of mind. However, most research of the emotion of videos has assumed that an emotion is unified at any given moment.

In the present paper, we propose a new temporally changed emotional analyzer that functions as a probabilistic estimator of multiple emotions evoked by videos. The goal of this study is to generate sequentially changing emotional responses from a video clip. This paper is organized as follows: First, a psychophysical experiment to investigate the evoked emotions by each scene is described. Second, the proposed system is introduced with mid-level visual features, an estimation model, and a performance test. Finally, the conclusions are summarized and future tasks are proposed.

Method

Data Set

To investigate the evoked emotion of each video clip, stimuli were taken from following movies and TV series: *Amélie of Montmartre (2001), Artificial Intelligence: AI (2001), Curse of the Golden Flower (2006), The Amazing Spider-Man (2012), Wuthering Heights (2012), Friends Season8: The One Where Rachel Tells Ross (2001),* and *CSI: Miami Season8 Episode4: In Plane Sight (2009).* The movies and TV series were selected non-intentionally. Each was decomposed to a set of video scenes, the emotional valence of which varied. Thirty-three video clips were selected in order to not exceed one hour per experiment per person. The average run time of the clips was 81 seconds. Each video clip had one main event that occurred in one location. We preferred that each video clip evoke one kind of emotion.

Evoked Emotions

To examine the emotions evoked, we used emotional models instead of Canini's 3D emotional identity space (Canini et al., 2009). Ekman's model is defined by 6 basic facial emotional expressions: anger, surprise, disgust, sadness, happiness, and fear (Ekman & Friesen, 1978). However, this model may not be applicable for video-evoked emotions because it contains more negative emotions than positive ones: it has 4 negative emotions (anger, disgust, sadness, and fear), one neutral emotion (surprise) and one positive emotion (happiness). Thus, the model needs to be balanced. Moreover, the model did not include some key emotions that viewers experience while watching videos. Gross and Levenson (1995) modified the 6-emotion model to an 8-emotion model. The 8-emotion model comprises the following emotions: amusement, anger, contentment, disgust, fear, neutral, sadness and surprise. Unfortunately, this model is also not balanced.

We designed a 12-emotion model that includes Gross and Levenson's (1995) 8-emotion model with 4 emotions (humor, romance, tension, and suspicion). Our emotional space is balanced in that there are 4 positive emotions (humor, romance, contentment, and joy/pleasure), 4 blended emotions (suspicion, tension, surprise, and neutral), and 4 negative emotions (anger, disgust, sadness, and fear).

Procedure

20 people participated in the experiment. At the beginning of the test, the 12 emotions were described (Fig. 1) to the participants. The video clips were shown to the participants on a TV monitor. The presenting order of videos was randomized to minimize potential error. After watching each video clip, the participants were asked to choose 1 of the 12 emotions that best represented the emotion evoked by the video clip.

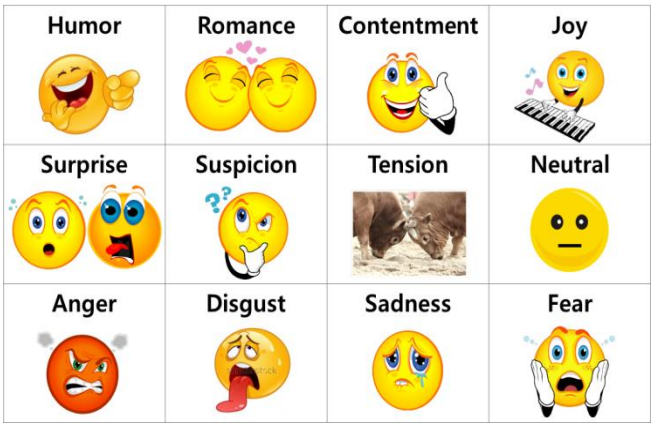

Figure 1 : The names and illustrations of the 12 emotions used in this experiment.

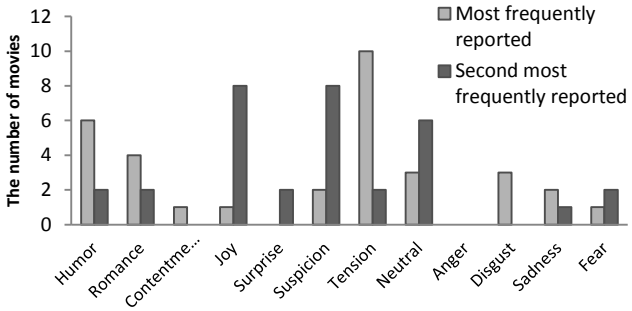

Figure 2 : Emotions evoked by the 33 video clips.

Results

Responses of Evoked Emotion

The participants' responses are summarized by a 12-emotion population for each video clip. The results of the evoked emotional responses for the 33 video clips are illustrated in Figure 2. The figure shows the frequency of the most commonly reported and the second most commonly reported emotion per video clip. Some emotions were not reported often; this may generate some noise in our model. Gross and Levenson (1995) reported that contentment, anger, and fear are more difficult to be evoked by movies than are other emotions. Notably, anger was not evoked by the video clips.

Statistical Analysis

To determine whether the response data varied according to the video clip, we analyzed the data statistically using SPSS 1.9. First, we tested independence of the participants' responses for all of the test video clips, through a cross-tabulation analysis using chi-square tests. The relationship between the evoked emotion and the video clip presented was significant ($p < 0.001$).

Second, the independence of participants' evoked emotions was evaluated in the several extracted video clips from the same movie or TV series. The video clips were composed of 3 clips of *AI*, 5 clips of *CSI*, 4 clips of *Wuthering Heights*, 4 clips of *Amelie*, 4 clips of *Friends*, and 12 clips of *Spider-Man*. The evoked emotion was significantly related to the video clip ($p < 0.001$). The results provided in Tables 1 and 2 are examples of our statistical analysis applied to the data collected in response to the 3 clips of *AI*. While the *AI* clips evoked all of the 12 emotions, the other clips evoked as 3 or 4 emotions.

Table 1 : The results of the cross-tabulation analysis for the *AI* video clips.

Count	Emotion												Total
	Hum	Rom	Con	Joy	Sur	Sus	Ten	Neu	Ang	Dis	Sad	Fea	
AI_1	0	0	0	1	0	2	13	3	0	1	0	0	20
AI_2	1	0	0	0	3	0	0	0	3	9	2	2	20
AI_3	0	4	9	3	0	0	0	0	0	0	4	0	20
Total	1	4	9	4	3	2	13	3	3	10	6	2	60

Table 2 : The results of the chi-square test for the data collected in response to the *AI* video clips.

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	102.100	22	.000
Likelihood Ratio	113.195	22	.000
Linear-by-Linear Association	6.362	1	.012
N of Valid Cases	60		

Estimation Model of Evoked Emotions

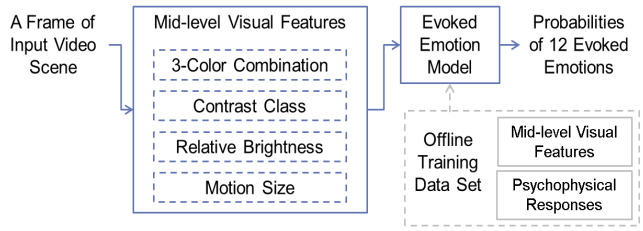

Figure 3 : A block diagram of our proposed estimation system.

Do computers mimic human emotions such as the emotions evoked by the video clips? To answer this question, we proposed a movie-evoked emotion estimator by introducing mid-level visual features. The mid-level visual features were composed by referring to color emotional theory with a 3-color combination and emotionally correlated general visual features: contrast by histogram distribution, relative brightness, and motion size. To create the evoked-emotion estimator, we examined the relationship between humans' psychophysical responses and the mid-level visual features extracted from the video clips. The emotion model was trained by supervised learning with humans' psychophysical responses and their features. The proposed approach generated multiple emotional states of the video contents sequentially and is illustrated in Figure 3.

Visual Feature Extraction

Psychologists have investigated the emotion-eliciting properties of industrial media (Gross & Levenson, 1995; Pos & Armytage, 2007; Kobayashi, 1981). In particular, Gross and Levenson (1995) focused on eliciting the emotion of films. They determined that each film could evoke 8 different emotions from the viewers. The results show that video clips can be categorized by the different emotions that they evoke from the viewers. However, they did not attempt to draw relationships between the visual cues and the elicited emotions. The present study focused on some mid-level visual cues and evoked emotions from some video clips. The mid-level visual cues are reorganized by an analysis of low-level features. Moreover, the mid-level cues were differentiated from low-level cues. The mid-level cues were extracted by color, contrast, brightness, and motion.

Color Color is known to correlate strongly with psychological constructs. Many studies describe the relationships between these variables (Pos & Armytage, 2007; Kobayashi, 1981; Solli & Lenz, 2010), but to date, correlations in movie settings have not been studied. Pos and Armytage (2007) investigated the relationships between emotions, facial expressions, and colors. Kobayashi (1981) matched 1170 3-color combinations to 180 adjective words describing emotional appearance. Solli and Lenz (2010)

transformed and classified natural images on the basis of Kobayashi's list of emotional words. Kobayashi's color scale is useful for industrial design. Unfortunately, it is not appropriate to apply the 180 emotional words (i.e, warm, cold, luxury, etc.) to movies. Pos and Armytage (2007) studied the relationships between 3-color combinations and Ekman's 6-emotional facial expressions. The 3-color combination might connote simple feelings (e.g., warm, cold) and emotions (e.g., happy, sad).

We extracted the color distribution of a video clip to estimate evoked emotions. We transformed the colors of the input video clips into Kobayashi's 130-color scale Hue and Tone System. The transformation helped to reduce the complexity of the analysis ($2^{8\times3} \rightarrow 130$) and to classify the input colors into emotionally meaningful representative colors. An input pixel color (RGB) is converted to the Hue and Tone value (HT) as shown in Equation 1.

$$RGBtoHT(x) = \arg\min_{i} \{RGB\_Dist(x, HT[i])\} \qquad (1)$$

where x is 1 of the $2^{8\times3}$ colors used as an input color, *HT*[i] is 1 of the 130 Hue and Tone colors ($1 \leq i \leq 130$), and *RGB_Dist(x,y)* is a distance measuring method such as Euclidean distance on RGB space. The HT with the minimum distance was selected as the HT value of the pixel. Then, we extracted 3 colors used most frequently in the frame. The 3 dominant colors were accumulated for all of the frames of a video clip as a probability distribution of hue and tone colors. The entire process is illustrated in Figure 4.

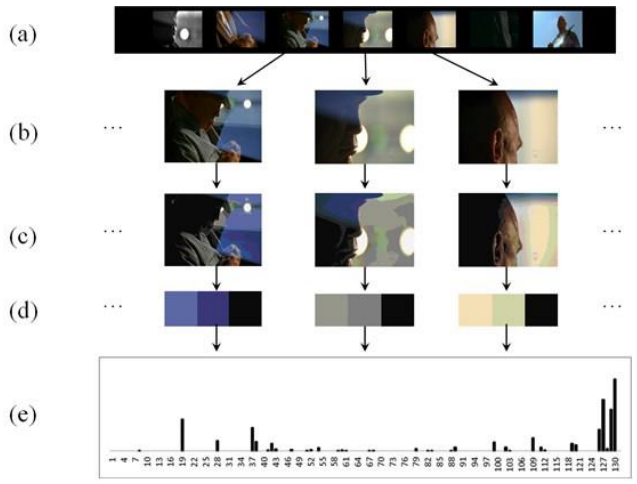

Figure 4 : Our color extracting method: (a) a video stream; (b) RGB images of each of the frames; (c) converted HT images of each of the frames; (d) three dominant colors of each of the frames; and (e) a probability distribution of HT for all of the frames.

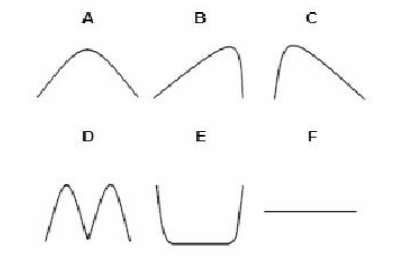

Figure 5 : Contrast classification method (Kim, Choe, & Lee, 2006).

Contrast Except for the color feature, we can extract many features from visual information. Above all, a high-contrast scene may evoke a different emotion than a low- or mild-contrast scene. We used a contrast-categorization-method based histogram analysis of image intensity (Kim, Choe, & Lee, 2006). The categorization performance of the method has already been proven by a previous human psychophysical experiment. The method assigns the histogram distribution of every image to 1 of 5 representative contrasts. We merged class **D** and class **E** (see Figure 5) into 1 category since they yield similar contrast.

Brightness Image brightness is 1 of the important cues serving as a connotative feature of videos. Low brightness may not be used by directors to express a hopeful scene. However, it would be inappropriate to use absolute brightness as an estimating tool for eliciting emotion, because the overall brightness of a video depends on the brightness characteristics of the video camera or the director's preference. Consequently, we extracted the relative brightness ($\tilde{Y}$) to the average brightness of the entire video as described in Equation 2. Y_f *(i)* is an average brightness of all of the pixels in the *i*th frame, *n* is the total number of frames for a scene, and *m* is the total number of frames for an entire video.

$$\tilde{Y} = \frac{\sum_{i=0}^{n} Y_f(i)}{n} - \frac{\sum_{j=0}^{m} Y_f(j)}{m} \qquad (2)$$

Motion Motion is an important factor in the evaluation of how dynamic scene is. We used a scale-invariant feature transform (SIFT, Lowe, 2004) to obtain motion information. SIFT features are extracted from the video frames and their trajectory is evaluated to estimate inter-frame motion (Lowe, 2004). In this study considers the distance of their trajectory was considered as the motion size. Then, the motion feature defined the largest motion size of each inter-frame.

Emotion Estimation

To estimate emotions evoked in participants while watching videos, we used an evoked-emotion model. The psychophysical experiment was used to develop the emotion model and the multiple emotional responses were the

training data. Then, the emotion model was learned by a supervised learning method using the psychophysical responses.

Training Data As mentioned above, we extracted a motion feature from inter-frame data and 3 other features from intra-frame data. We needed some representative values of the features to train the emotion model. Thus, each feature was analyzed in every frame of a clip, and all feature data were summarized during 1-second intervals. All video clips were standardized to the same pixel resolution and frame rate. Every clip was composed of 24 frames per a second. The color feature was encoded by using a 130-variable combination. The variable was calculated by normalization (between 0 and 1) of the accumulated HT probability distribution for all of the frames. The contrast feature was encoded by using a 5-variable combination and the variable was presented by the normalization of the accumulated probability distribution for contrast categories of all of the frames. Brightness and motion features were calculated by averaging them over all of the frames.

A training set of supervised learning consisted of an input feature vector and an output target vector. The target vector was designed as 6 probability values. The probability values were such that five emotions (contentment, surprise, anger, sadness, and fear) were filtered out to prevent mis-learning by null data (see Fig. 2). Moreover, "neutral" is a broad emotional term with diverse meanings and, therefore, it was also filtered out. Owing to the removal of the emotions, 6 clips in which one of the removed emotions predominated had to be excluded.

Estimation Model The emotions evoked in the participants were the target data used to train the model. Classification and Regression Trees (CART, Breiman, et al., 1984), a kind of decision tree learning, was selected as the learning method to classify the evoked emotion from a video. One advantage of CART is that it can consider misclassification costs in the tree induction, using handling numerical-valued attributes. The model learns a probabilistic response for each emotion and the extracted visual features. The participants' responses are set as the independent variables and mid-level visual features serve as the dependent variables of the model.

Estimation Performance Finally, we conducted supervised learning with CART. The model outputs probability values (between 0 and 1) for every 6 emotions for an input video. We evaluated the performance with a between-clip cross-validation of 27 video clips.

A summary of our estimation performance is shown in Table 3. For the sake of convenience, the accuracy of performance was calculated by using an emotion with the estimated maximum probability for each clip. That is, each of the 27 videos had a predicted emotion with a majority probability. The predicted emotions were compared with the most common emotion of the target data; the correction rate was 56%. In Table 3, 2nd emotion accuracy refers to the percentage of times that either the most common or the second-most-common emotions was the predicted emotion.

Table 3: Details of the accuracy of the evoked emotion model.

	1st emotion	2nd emotion	3rd emotion
Prediction accuracy	56%	63%	70%

Figure 6 shows a partial emotional profile of *The Amazing Spider-Man* using our model. The profile shows that the movie does not have 1 evoked emotion; instead, it has 3 or more evoked emotions at the same time. In addition, each emotional state is shown as a probability value. The emotional state exhibits variations that are similar to the original characteristics of the movie. For instance, the middle of the movie has many scenes with a heroine and a hero that induce a romantic mood, and the latter part has many episodes in which the hero challenges a powerful villain. Further, the dominant emotion of each second from Figure 6 is illustrated in Figure 7. The pie chart might help one to search a movie for some of the most important information.

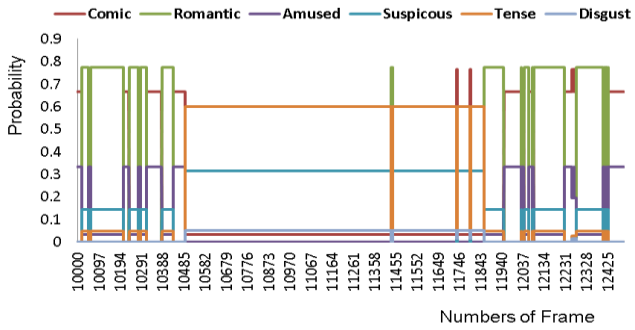

Figure 6 : A partial emotional profile of *The Amazing Spider-Man* using the proposed method.

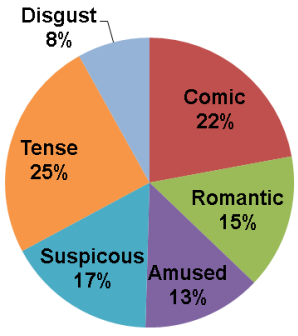

Figure 7 : The emotional composition of *The Amazing Spider-Man.*

Discussion

Previous attempts to characterize the emotions of videos have used a single emotion, as a representative emotion for an entire video sequence (Gross & Levenson, 1995; Bailenson et al., 2008; Canini et al., 2009). However, a person's emotions vary every moment (Winter & Kuiper, 1997).

To take these individual differences into consideration, the present paper proposes a probabilistic model that can estimate multiple emotions evoked by a video over time. The proposed method involves the automatic labeling of videos with the emotions that were evoked and the duration of the evoked emotions. The present study attempted to characterize the evoked emotions by using mid-level features from the video frames, such as dominant colors, contrast, brightness levels, and motion quantification. The characterization was derived from a previous psychophysical experiment using human participants. A classical machine learning methodology was applied in order to build and test the model of the emotional categories targeted.

The present study provides 2 significant advances. First, it is the first to propose a new paradigm for video retrieval, using probabilistic multiple-evoked emotions. This approach may be used to construct an emotion-based video retrieval system, a video recommendation system, or an emotional treatment system using a video. It may also help a machine to generate a new video profile that automatically describes sequentially changing emotions. Second, this study provides a technically new approach. That is, we used emotionally meaningful mid-level visual features and we modeled them to estimate multiple-evoked emotional states from videos.

However, there are limitations to the present research. The experimental video clips were limited and participants made non-various responses (Figure. 2). Having participants choose only a single emotional response in the experiment might lead to inadequate data for generation of a probabilistic emotional estimator. One of the interesting findings was that there were no anger responses. Mikels, et al. (2005) reported the same result with still images. The authors concluded that anger is very difficult to elicit with the passive viewing conditions of static images. Our experiment used some video clips that were a few minutes long. One explanation for the finding of no anger responses is that the video clips were not long enough to evoke the emotion.

Thus, 3 tasks are left for future research to implement our approach in a real system. First, more video clips are required to teach the various emotions to the evoked emotion model. Second, a larger sample is required to get a robust trained model. Third, other machine learning methods should be tried to find an optimal solution.

Acknowledgments

This work was supported by the National Research Foundation (2012-0005643, Videome) grant funded by the Korea government (MEST) and was supported in part by the Industrial Strategic Technology Development Program (10044009) funded by the Korea government (MKE).

References

Bailenson, J.N. et al. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies*, 66, 303-317.

Breiman, L. et al. (1984). *Classification and Regression Trees*, New York, Chapman & Hall (Wadsworth, Inc.).

Canini, L. et al. (2009). Emotional identity of movies. *IEEE International Conference on Image Processing (ICIP*). 1821-1824.

Ekman, P. & Friesen, W. V. (1978). *Manual for facial action coding system.* Palo Alto: Consulting Psychologists Press.

Gross, J.J., & Levenson, R.W. (1995). Emotion elicitation using films. *Cognition and Emotion*, 9, 1, 87-108.

Kobayashi, S. (1981). Aim and method of the color image scale. *Color Research & Application*, 6, 2, 93-107.

Kim, I., Choe, W., & Lee, S.D. (2006). Psychophysical measurement for perceptual image brightness enhancement based on image classification. *Proceedings of the SPIE*, 6057, 425-431.

Li, S., Zhang, Y.J., & Tan, H.C. (2010). Discovering Latent Semantic Factors for Emotional Picture Categorization. *IEEE International Conference on Image Processing (ICIP*). 1065-1068.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2, 91-110.

Mikels, J.A. et al. (2005). Emotional category data on images from the International Affective Picture System, *Behav. Res. Methods* 37, 626–630.

Pos, O., & Armytage, P. G. (2007). Facial Expressions, Colours and Basic Emotions. *Colour: Design & Creativity* 1(1) 2, 1–20.

Solli, M., & Lensz, R. (2010). Color semantics for image indexing. *Proceedings of 5th European Conference on Colour in Graphics, Imaging, and Vision*, 353-358.

Yanulevskaya, V. et al. (2008). Emotional valence categorization using holistic image features. *IEEE International Conference on Image Processing (ICIP*). 101-104.

Wang, W., & He, Q. (2008). A survey on emotional semantic image retrieval. *IEEE International Conference on Image Processing (ICIP).* 117-120.

Winter, K. A., & Kuiper, N. A. (1997). Individual differences in the experience of emotions. *Clinical Psychology Review*, 17(7), 791– 821.