PHOTOSENSE: MAKE SENSE OF YOUR PHOTOS WITH ENRICHED HARMONIC MUSIC VIA EMOTION ASSOCIATION

Ja-Hwung Su [†], *Ming-Hua Hsieh* [†], *Tao Mei* [‡], *Vincent S. Tseng* [†]

[†] National Cheng Kung University, Taiwan
[‡] Microsoft Research Asia, Beijing, P. R. China

ABSTRACT

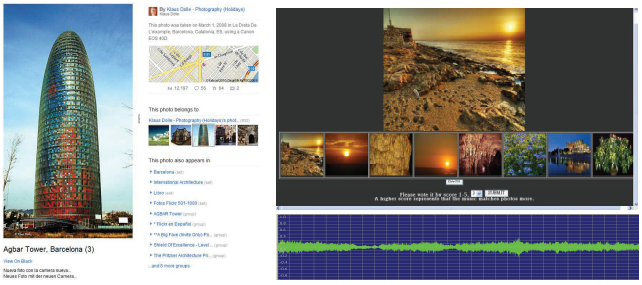
This paper proposes a novel audiovisual presentation system, called PhotoSense, to enrich photo navigation experience by associating emotionally harmonic music with a given photo collection. Different from many conventional photo visualization systems which predominantly focus on the visual elements for presentation, we explore both visual and aural perspectives which can enhance the browsing experience from each other. This is achieved by building an emotion space shared by visual and aural domains, and a set of emotion classifiers which can associate each visual and aural element with this space. Furthermore, we design a sequence matching algorithm to associate a set of music with a photo collection by maximizing similarity in the emotion space. PhotoSense represents one of the first mash-up applications which build a natural connection between the ever increasing personal photo collections on the Web and music-sharing sites. Experiments show that PhotoSense provides better browsing experience for photo collections.

Index Terms— Multimedia presentation, affective content analysis, emotion.

1. INTRODUCTION

The prevalence of image capture devices and the advent of media-sharing services such as Flickr have drastically increased the volume of community-contributed images. People are spending much more time to browse these images, as well as using these images to share their experiences, than ever before. Most existing image-oriented sites provide a straightforward management tool to allow users browse their interested photos along a single dimension—visualization, while another important sense—audio—is usually neglected. It has been known that people browse photos with emotion. In other ways, the photo-based storytelling always expresses some emotion, e.g., happiness or sadness.

In this paper, we are investigating if we can provide an enriched photo browsing experience by accompanying some similar background music with a given photo collection, so that people can more easily understand the underlying story and have a better browsing experience. The accompany music is emotionally harmonic with the photos, so that this rich presentation can provide a better atmosphere (with both appropriate visual and aural senses) for browsing. For example, if a photo album is about a personal trip to Barcelona, then when a user is browsing this album, the local famous "flamenco music" will be automatically triggered with these photos.



(a) Flickr (b) PhotoSense

Fig. 1. Examples of photo browsing systems. Different from traditional browsing services which only focus on photo visualization, PhotoSense provides both visual and aural (accompany music with harmonic emotion) aspects for enriching user experience.

Figure 1 shows the example of typical photo browsing system (i.e., Flickr) and our proposed system in this paper (called PhotoSense). We can observe that while Flickr only displays photos linearly, PhotoSense is able to automatically recommend background music which is consistent with the visual content in terms of emotion. We know that the photos in this collection look sad and thus associate some music with "sad" emotion with it. When a user is browsing this photo collection, he may use at least two types of sensation to enjoy the story conveyed by the joint media presentation. The similar idea can be found in movies. A movie with background music can get user much more involved in the atmosphere than that without background music. The affective sensation enriched by the music can futher enhance the browsing atmosphere. This inspired us to design an audiovisual presentation system by connecting photo to music through a common emotion space. Although there exists research on modeling

978-1-61284-350-6/11/$26.00 ©2011 IEEE

the relationship between visual and aural domains [1] [2] [3] [4], this paper proposes to use emotion as the intermediate linkage between them which is a new perspective.

To further investigate how much users would like this audiovisual presentation style, we invited 25 volunteers to participate with a feasibility user study. As a result, among these volunteers, most of them like this presentation style except for two. The main reason that the users like this presentation is that, it can really enrich browsing experience by the associated music with harmonic emotion.

In paper, we report our preliminary effort on building a new audiovisual presentation system, called PhotoSense, which is able to associate a background music with a given photo collection. This association is designed to minimizing the difference between visual and aural emotion in a common space. First, photos and music are mapped into an emotion space via classifiers. Then, a sequence matching algorithm is designed to associate photos and music in this space. PhotoSense connects music and photo by matching emotion styles. Through the connection of music and photos, users can browse the photos with affective background music. The main contributions of this paper can be summarized as follows:

- We propose an audiovisual presentation system that is able to enrich user experience of photo browsing.
- We propose to link photo and music via emotion, which is a new perspective and significantly different from related research.
- We build a common emotion space via classifiers on visual and aural domains.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the details of PhotoSense system. Empirical evaluations are given in Section 4, followed by the conclusions in Section 5.

2. RELATED WORK

We review previous research on the association between visual and aural forms from three aspects: 1) given a music, associating images/videos with it [1] [5] [6] [7]; 2) giving a collection of images, associating music with them [8] [3] [4]; and 3) emotion analysis from media [9] [2] [10]. We also compare our proposed PhotoSense with the related research.

The most emerging multimedia application integrating visual and aural elements is photo browsing which embeds adaptive music to a collection of photos [1] [6] [7]. The Tiling slideshow system integrates visual and musical forms, which can display photos in a tie-like manner and coordinate with the pace of background music [1]. In [6], the authors proposed an approach of music generation from image based on visual-auditory association. The emotional linkage between

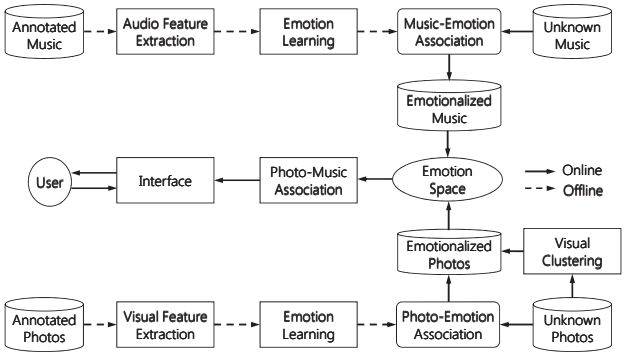


Fig. 2. System overview of PhotoSense.

image and music was established through mapping the saturation and brightness of image to the key and tempo of music. For example, the image with low brightness and saturation has a lower key, while the image with low brightness but high saturation has a stronger temp. Although the establishment of the emotional linkage between image and music was intuitive in [6] [7], the link based on the characteristic of color, temp, and key may not be stable as it neglects users' feelings about music and image. The resolution of what kind of emotion an image or music brought to user would facilitate such emotion linkage. However, how to determine user's feeling towards image and music remains a challenge. Hua *et al.* presented an automated home video editing system, in which users can give a piece of incidental music, and then the system extracts a series of video segments satisfying certain editing rules [5].

Kabisch *et al.* proposed a system to create accompanying music based on a given image [8]. Wu and Li presented several algorithms to generate music segments from image features based on the mapping from global image attributes to musical attributes [3] [4].

In [10], Yoo utilized the human evaluation of color pattern on 13 emotion scales to design query with emotion terms for an image database. Wang *et al.* [2] identified an orthogonal three-dimensional emotion space and designed three emotion factors for artwork. In [9], the authors used standard RGB histogram and psychophysics-based and color-emotion-based features to characterize the global appearance of image.

We can observe that most existing research focused on either photo/music generation based on visual-auditory association or the emotion analysis form photo/music. In this work, we utilize the emotion patterns learned from music/image feature to design the adaptive background music.

3. PHOTOSENSE

3.1. Overview

Research on multimedia has focused on understanding content by associating the content with semantic concepts [11] [6]. However, there exists other valuable information, such as

affective senses. In fact, the affective senses hidden in multimedia content are very helpful to presentation. How to determine the visual and musical emotion and associate these two different domains in a natural way is the main issue in this paper. The PhotoSense system is characterized by discovering implicit affective content to support a natural and enriched audiovisual presentation. As shown in Figure 2, the system can be decomposed into two parts: offline learning and online visual-aural presentation.

- Offline Learning Part. This part refers to the modeling of visual and audio affective emotions from annotated photos and music, respectively. We have built an emotion lexicon shared by both visual and aural domains, as well as several classifiers based on low-level features for identifying the emotion categories for each photo and music.
- Online Presentation Part. Given a new photo or music, we will classify each into emotion space. Then, using a photo-music association module based on the space, music is associated with the given photo collection.

3.2. Associating Photos with Emotions

In a photo presentation system, visual analysis is important since it can influence user's browsing atmosphere by emotion hidden in the content. For example, a photo taken in a sunny day brings more pleasure emotion that that in a rainy day. In the previous works [9] [12], *pleasure*, *arousal*, and *dominance* are widely adopted as the affective senses to present human emotion. However, it is difficult for an average user to define the relation between visual content and dominance. Thus, we only adopt two emotions in this paper, including *pleasure* and *arousal*. In general, pleasure stands for the level of happiness, while arousal for the level of activation. From the viewpoints of pleasure and arousal, it is easy for the user to define how she/he feels while looking at a photo. The pleasure and arousal are divided into three levels ranging at a scale of 1-3. Therefore, the problem can be reduced to a classification. The notion behind classification is that, photos with visual features are projected onto the level of the affective sense by classifiers. Hence, a photo can be classified as one of three levels for pleasure and arousal.

In the followings, we describe how to extract the related visual features. Each photo is divided into 5×5 grids. Brightness (b) and saturation (s) are extracted to describe a grid. For each grid, the average (u), standard deviation (σ) and skewness (w) for both brightness and saturation are calculated, in which w is defined as follows.

$$w_b = \frac{1}{n}\Big(\sum_{i=1}^{n}(b_i - u)^3\Big)^{\frac{1}{3}}, w_s = \frac{1}{n}\Big(\sum_{i=1}^{n}(s_i - u)^3\Big)^{\frac{1}{3}} \quad (1)$$

where n denotes the size of photo, w_b and w_s denote the skewness for brightness and saturation, respectively. As a result, six features can be extracted from a grid. Then, the feature dimension for each photo is 150 (6×25). Two classifiers, namely visual pleasure classifier and visual arousal classifier, are modeled by a set of training photos for pleasure and arousal, respectively.

3.3. Associating Music with Emotions

For each music, 38 frames within one second are extracted and each frame is represented by 576 Modified Discrete Cosine Transform (MDCT) coefficients within 26ms. In this paper, 36 important coefficients instead of 576 coefficients are further selected to reduce the computational costs. Three music features are extracted for each music, i.e., Average Energy Difference (AED), Average Intensity (AI), and Average of Low-Energy Frame Quantity (ALFQ). Assume that a music is decomposed into a set of frames $F = \{f_1, f_2, \ldots\}$, and the i-th frame consists of 36 MDCT coefficients $f_i = \{m_1^i, m_2^i, \ldots, m_{36}^i\}$. Three music features are given by

$$\begin{aligned} AED &= \frac{1}{|F|} \cdot \frac{\sum_{j=1}^{|F|}\sum_{i=1}^{36}|m_i^j - m_{i-1}^j|}{36} \\ AI &= \frac{1}{|F|}\sum_{j=1}^{|F|}\sum_{i=1}^{36}\left(m_i^j\right)^2 \\ ALFQ &= \frac{1}{|F|}\sum_{j=1}^{|F|} I^j \end{aligned} \quad (2)$$

where $I^j = 1$, if $Avg^j \leqslant 0.5 \cdot \overline{Avg}$; otherwise, $I^j = 0$. $Avg^j = \sqrt{\frac{\sum_i m_i^j}{36}}$, $\overline{Avg} = \frac{1}{|F|}\sum_j Avg^j$. The major intention for our extraction strategy is to present the user's emotion by music intensity. Finally, three features extracted from music are adopted to establish the music-emotion classifier for pleasure and arousal, respectively, namely musical pleasure classifier and musical arousal classifier.

3.4. Presentation by Photo-Music Association

After model visual and aural emotion, we can construct an audiovisual presentation system to create a harmonic atmosphere by emotion. The motivation is that— "if one day you look at a set of pictures in a good mood, you perhaps desire some light music." To this end, a key issue is how to connect music and photos by affective senses.

For constructing an audiovisual system, the presentation style has to be determined. Here the scenario is: a user's photos are grouped into k clusters by saturation and brightness. As user visits this system, the system selects q clusters randomly. For each cluster, one photo is randomly selected as a start. Once a start is picked, photos within this cluster are presented one by one in a slideshow. To be consistent with the presentation style, the association from music to photos can be further divided into three issues: 1) the duration of music, 2) clustering of photos, and 3) sequence matching of music

Algorithm 1 Association of photo and music collection.

Input: A set of photos and music with emotion.
Output: A display list of the relevant photos and music.

1: **for** each photo **do**
2: extract the brightness and saturation
3: cluster the photos by brightness and saturation by K-means
4: **end for**
5: **for** each cluster **do**
6: display a photo randomly;
7: **end for**
8: **if** photo p_i^j in the j-th cluster is picked by the user **then**
9: calculate the similarities S_p between p_i^j and the photos in the j-th clusters using the emotion features;
10: generate the display list L_p by sorting the photos by S_p;
11: calculate the similarities S_m between p_i^j and the music in the database using the emotion features;
12: generate the display list L_m by sorting the music by S_m;
13: **end if**
14: display the photos and music simultaneously by S_p and S_m;

and photos. Algorithm 1 gives the steps for associating music with a photo collection.

Duration of music. For an audiovisual presentation system, music has to be changed along with photo's change. If the browsing time for a photo is 6 sec, then music has to be changed very frequently, which does not make sense. On the contrary, it is not a good design to present a large number of photos by only one music. To deal with this problem, music duration is fixed to 30 sec. That is, music is changed per five photos.

Clustering of photos (as shown in Algorithm 1, line 1-4). For a user, it is hard to browse a large number of photos. It motivates us to group the photos into a set of clusters by brightness and saturation. From the viewpoint of affective sense, a cluster consists of the similar affective photos through visual-based clustering. In this stage, the clustering is performed by K-means.

Sequence matching of music and photos (as shown in Algorithm 1, line 8-13). As mentioned in Sections 3.2 and 3.3, each photo and music can be classified by visual pleasure, visual arousal, musical pleasure and musical arousal classifiers, respectively. Thus, a photo can be presented 4-dimensional visual affective features by {*pleasure level*, *pleasure confidence*, *arousal level*, *arousal confidence*}, where the pleasure and arousal level indicate the set of {1, 2, 3}, and confidence is the result derived by the referred classifiers. Music can be also presented a set of musical affective features by {*pleasure level*, *pleasure confidence*, *arousal level*, *arousal confidence*}.

Through the affective classification mentioned above, the affective senses of photos and music can be detected to support the mapping of photos and music. In our design, once the i-th photo p_i^j in the j-th cluster is picked up by the user, the mapping procedure is then triggered. The pleasure and arousal levels are employed as the affective feature vectors for photos and music. Next, according to the affective features, the similarities between p_i^j and music are derived by computing Euclidean distances. Besides, from content-based image retrieval viewpoint, the picked photo can be regarded as a query photo. As a result, the similarities between p_i^j and the photos in the j-th cluster need to be calculated to sort the photos. Hence the photos and music are bridged successfully. Finally, sorted photos and music are presented simultaneously.

4. EXPERIMENTS AND EVALUATIONS

4.1. Data and Methodologies

The experimental data consists of the collections of photos and music. We collected 3,000 photos from several image search engines, such as Google, Yahoo, Flicker, and so on. The queries include "natural categories of autumn," "flower," "mountain," "beach," "river," "grass" and "forest." 2,100 photos were selected as the training data and the remaining as testing. We grouped the testing images into 50 clusters. We also collected 1,000 music from Amazon, which covered six genres including *Classic Rock*, *Classic*, *Jazz*, *Latin Music*, *Opera Vocal*, and *Rock*. The duration for each piece is 30 sec. In the experiments, we adopted different classifiers, including Support Vector Machine (SVM), KNN ($K = 15$), Naive Bayes (NB), and Decision Tree (DT). We invited six subjects with diverse background to tag the photos and the music with their emotions. Each photo and music was annotated by more than three users and the final annotation is determined by majority voting.

4.2. Evaluations

4.2.1. Objective evaluation of associating photo and emotion

In fact, it is still a challenging issue of how much the affective sense is related to visual features. The major goal of this experimental evaluation is to understand the relation between photos and affective senses, such as pleasure and arousal. Figure 3 shows the evaluation results of associating photos and affective senses. The results deliver some interesting aspects. First, the precisions of different classifiers are different. The effectiveness of SVM, KNN, DT are pretty close for arousal, but higher than NB. Four classifiers perform similarly for pleasure, but NB is slightly worse than the other three. Overall NB is the worst whatever for pleasure and arousal. Second, arousal is more sensitive than pleasure and the precision reaches 90%. The potential reason is that, arousal is more easily defined for the users than pleasure while manually annotating the photos. On the whole, SVM and KNN are the good choices for detecting visual affective senses.

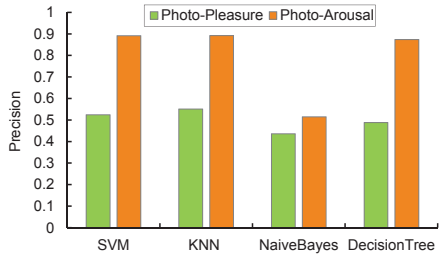


Fig. 3. Performance of visual pleasure and arousal classifiers.

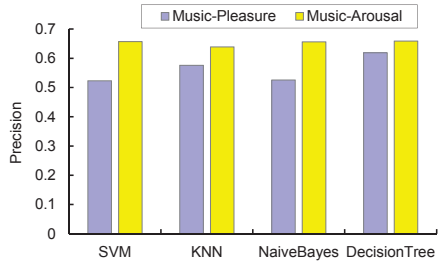


Fig. 4. Performance of music pleasure and arousal classifiers.

4.2.2. Objective evaluation of associating music and emotion

In addition to the evaluation of associating with photos and affective senses, another issue we are interested in is the evaluation of associating music and affective senses. Similar to the goal of the above evaluation, this experiment can help us clarify the implicit affective senses hidden in music. Figure 4 shows the resulting precisions for examining different musical affective classifiers. From Figure 4, we can obtain that: First, the precisions of four classifiers are very close for pleasure and arousal. DT performs slightly better than the others for pleasure. Second, quite similar to visual evaluations, arousal is also more easily to predict than pleasure. Third, the precision difference between pleasure and arousal for music is smaller than that for photos. Moreover, the precision for musical pleasure is higher than that for visual pleasure, but the result is contrary to that for arousal. It says that, musical arousal is more diverse than visual arousal. In contrast, pleasure is more consistent in music content than that in visual content. From psychology point of view, it is easier for humans to present her/his arousal felling on photos and her/his pleasure felling on music. Based on the above experimental evaluations, we adopted SVM as the visual and musical affective sense classifiers to further test our proposed audiovisual presentation system PhotoSense.

4.2.3. Subjective evaluations

In order to further evaluate the proposed audiovisual system, we conducted a photo presentation website by integrating visual and musical affective senses. Figure 5 is the screen snapshot of the home page of our proposed system. This system prototype is based on the assumption that, before presenting the photos, a number of photos have been grouped into 50 clusters by brightness and saturation. In Figure 5, a photo indicates a start selected from the related cluster. Once a photo is picked, the system displays the ranked photos in the related clusters with the mapping background music. Based on Figure 5, let us take the photos marked in red dot rectangles as examples whose concepts are Beach and Sunset. Figure 6 shows the resulting examples after the user picks the photos. Due to the limitation of paper presentation, we further visualize the mapping of photos and music using Figure 7. Figure 7(a) shows the mapping of low brightness-and-saturation photos and low intensity-and-energy music, while 7(b) shows the mapping of high brightness-and-saturation photos and high intensity-and-energy music. The detailed music information for Figure 7 is shown in Table 1 and 2.

In this experiment, we announced the system information

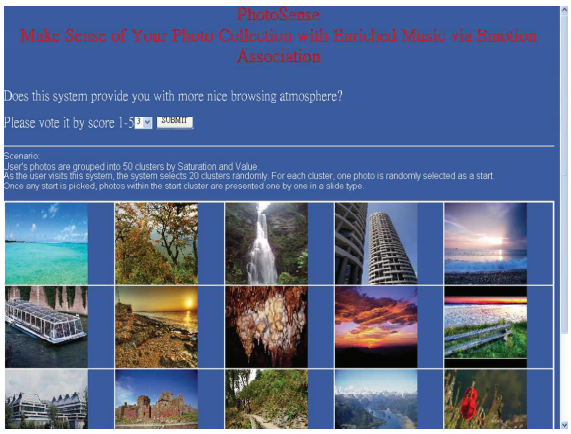
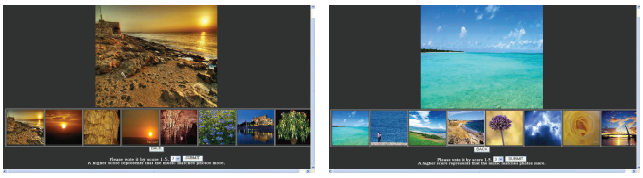


Fig. 5. PhotoSense demo site.



(a) (b)

Fig. 6. Visual examples after the user picks the photos.

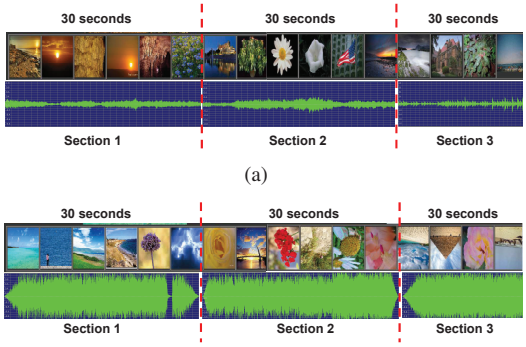


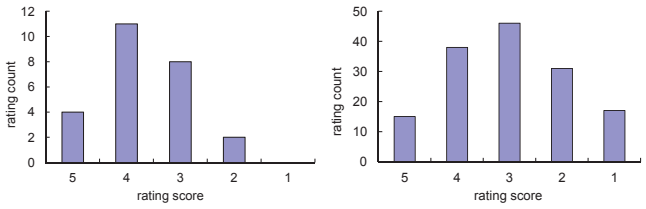
Fig. 7. Visual examples of associating photos and music.

Table 1. The detailed music info in Fig. 7(a).

	Music	Artist
Section 1	Amazing Grace	US Air Force Band
Section 2	God Of Our Father	US Marine Band
Section 3	Sheep May Safely Graze	Johann Sebastian Bach

Table 2. The detailed music info in Fig. 7(b).

	Music	Artist
Section 1	Afterlife	Avenged Sevenfold
Section 2	It's All Over	Three Days Grace
Section 3	I'm Sorry	Flyleaf



(a) the whole system (b) the mapping scheme

Fig. 8. Evaluations of the whole system and mapping scheme.

by email. There are around 25 volunteers participating in this system evaluation, including unknown volunteers. For each volunteer, she/he rated the system by answering the question "Does this system provide you with more nice browsing atmosphere?", which is scored from 1 to 5. To investigate the effectiveness, each volunteer was requested to pick at least 5 photos from the starts referred to Figure 5. For each picked, the volunteer had to rate the mapping scheme by score 1-5. Finally, we obtained 25 and 151 ratings for the whole system and the mapping scheme, respectively. Figures 8 shows the evaluation results for the whole system and the mapping scheme. From the previous recommendation literatures [11] [6], the user's preference can be classified into positive and negative. The preference is positive if the score is larger than 2. Otherwise, it is negative. Consequently, Figure 8(a) says that, most users like this idea for the audiovisual presentation system. That is, the positive rate for the system evaluation is 92%, while the average rating score is 3.68. The evaluation result of the mapping scheme is similar to that of the whole system. From Figure 8(b), we can obtain that, the positive rate for the mapping scheme evaluation is 67.3%, and the average rating score is 3.01. Overall, this system design can increase the browsing atmosphere, but it leaves some improvement rooms for associating photos and music by affective senses in terms of effectiveness. Here we believe that, it is a good start of audiovisual presentation system.

5. CONCLUSIONS AND FUTURE WORK

We have shown in this paper that a good media presentation system should be able to reflect and invoke emotion. To create a pleasant browsing atmosphere, we have presented a new audiovisual presentation system called PhotoSense, by extracting visual and musical emotions from photos and music, respectively. The main idea is to discover the relationship between emotion and media contents. Through the extracted emotion, photos and music are naturally connected. By embedding this kind of emotion linkage into the proposed system, browsing experience is significantly enriched.

Our future work may include: 1) in addition to pleasure and arousal, more emotion types will be considered; 2) in addition to the involved visual and audio features, other features will be investigated; and 3) context information, such as tags, location, emotion from tags, etc., can be integrated to enhance the discovery of emotion.

6. ACKNOWLEDGEMENT

This research was supported by National Science Council, Taiwan, R.O.C. under grant No. NSC99-2631-H-006-002 and NSC99-2218-E-006-001.

7. REFERENCES

[1] J.-C. Chen, W.-T. Chu, J.-H. Kuo, C.-Y. Weng, and J-.L. Wu, "Tiling slideshow," in *Proceedings of ACM Multimedia*, 2006.

[2] W.-N. Wang, Y.-L. Yu, and S.-M. Jiang, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, 2006, pp. 3534–3539.

[3] X. Y. Wu and Z.-N. Li, "A study of image-based music composition," in *Proc. of IEEE International Conference on Multimedia and Expo*, 2008, pp. 1345–1348.

[4] X. Y. Wu and Z.-N. Li, "Exploring visual-auditory associations for generating music from image," in *Proc. of ACM Multimedia*, 2008.

[5] X.-S. Hua, L. Lu, and H.-J. Zhang, "Optimization-based automated home video editing system," *IEEE Trans. on Circuit and Syst. for Video Tech.*, vol. 14, no. 5, pp. 572–583, 2004.

[6] J. H. Su, B. W. Wang, C. Y. Hsiao, and V. S. Tseng, "Personalized rough-set-based recommendation by integrating multiple contents and collaborative information," *Information Sciences*, vol. 180, pp. 113–131, 2010.

[7] J. H. Su, H. H. Yeh, P. S. Yu, and V. S. Tseng, "Music recommendation using content and context information mining," *IEEE Intelligent Systems*, vol. 25, no. 1, pp. 16–26, 2010.

[8] E. Kabisch, F. Kuester, and S. Penny, "Sonic panoramas: Experiments with interactive landscape image sonification," in *Proc. of International Conference on Augmented Tele-Existence*, 2005, pp. 156–163.

[9] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states," *Genetic Social and General Psychology Monographs*, vol. 121, no. 3, pp. 339–361, 1995.

[10] H.-W. Yoo, "Visual-based emotional descriptor and feedback mechanism for image retrieval," *Journal of Information Science and Engineering*, pp. 1205–1227, 2007.

[11] M. Soli and R. Lenz, "Emotion related structures in large image databases," in *Proc. of ACM international Conference on Image and Video Retrieval*, 2010.

[12] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective mtv analysis based on arousal and valence features," in *Proc. of IEEE International Conference on Multimedia and Expo*, 2008, pp. 1369–1372.