Emotion-based Music Visualization using Photos

Chin-Han Chen[1], Ming-Fang Weng[2], Shyh-Kang Jeng[1], and Yung-Yu Chuang[2]

[1] Department of Electrical Engineering,
[2] Department of Computer Science and Information Engineering,
National Taiwan University,
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617, Taiwan
b90030@csie.ntu.edu.tw,mfueng@cmlab.csie.ntu.edu.tw
skjeng@ew.ee.ntu.edu.tw,cyy@csie.ntu.edu.tw

Abstract. Music players for personal computers are often featured with music visualization by generating animated patterns according to the music's low-level features such as loudness and spectrum. This paper proposes an emotion-based music player which synchronizes visualization (photos) with music based on the emotions evoked by auditory stimulus of music and visual content of visualization. For emotion detection from photos, we collected 398 photos with their emotions annotated by 496 users through the web. With these annotations, a Bayesian classification method is proposed for automatic photo emotion detection. For emotion detection from music, we adopt an existing method. Finally, for composition of music and photos, in addition to matching high-level emotions, we also consider low-level feature harmony and temporal visual coherence. It is formulated as an optimization problem and solved by a greedy algorithm. Subjective evaluation shows emotion-based music visualization enriches users' listening experiences.

Key words: Emotion detection, Music visualization.

1 Introduction

Media such as music and photos bring us different *emotions*, from *sadness* to *joy*, depending on the content of media. The integration of different forms of media could evoke even more feelings and give a more touching presentation as long as they are synchronized in emotions. Most music players for personal computers, such as *Winamp* and the *Microsoft Media Player*, are featured with *music visualization* by generating animated imagery when playing music. Some simply display patterns irrelevant to the music content while elaborate ones present visual effects with coordination to the music's low-level features such as loudness and frequency spectrum. There exists other more sophisticated forms of music visualization, for example, man-made music videos. However, their production involves experts and requires a lot of manual efforts. On the other hand, photo slideshow is also often accompanied with music and photos are switched at beat time to enhance viewer's watching experience. For better composition of music and photos, this paper proposes a system to create emotion-based music

visualization using photos. By coordinating emotions in both auditory and visual contents, emotional expression is enhanced and user's listening experience is enriched. Same technique could also be used to create more touching photo slideshows.

Recently, a lot of work has been done for emotion detection from music based on acoustical feature analysis. Lu *et al.*adopted a hierarchical framework to detect musical moods [1]. Though good performance was achieved, taxonomy of emotion classification is quite restricted in that paper, only four categories. Wu and Jeng expanded the taxonomy of music emotion to eight categories [2]. On the contrary, there are very few papers on automatic photo emotion detection if any. In this paper, we used the same emotion taxonomy proposed by Wu and Jeng [2] for both music and photos. The identical taxonomy facilitates the emotion-based integration of music and photos. With such an emotion taxonomy, we collected a set of images and annotated their emotion categories. For automatic photo emotion detection, we propose a set of visual features that influence human visual emotion perception. A Bayesian classification framework using these visual features is proposed and leads to satisfied classification results for our application.

There are various aesthetical strategies for combining visual and auditory media empirically and many methods to measure the similarity of two media according to their content. Perceptually, tempo and timbre of music are often related to camera motions and colors in images. In Tiling Slideshow [3], photos are displayed in a tiled fashion with background music and photos are switched in synchronization with tempo of the music. In Hua *et al.*'s work [4], the tempos of music and video are extracted and matched to create music videos. hilippe Mulhem *et al.* [5] tried to give additional impacts by combining video and audio from the view of aesthetics. There is however not much connection between high-level notions of both media. In this paper, we propose a system to combine two media based on their emotions. User evaluation shows that emotion-based presentation of music and photos is more impressive and affecting. Specifically, this paper has the following contributions: (1) a collection of images with emotion annotations, (2) an automatic photo emotion detection algorithm and (3) a scheme for composition of music and photos.

The rest of this paper is organized as following. Section 2 gives an overview of our system. Section 3 introduces the method for automatic photo emotion detection. The algorithm for composition of photos and music is described in Section 4. Finally, Section 5 presents evaluation for results and Section 6 concludes the paper.

2 Overview

Figure 1 gives an overview of the proposed system. The inputs to our system is a song and a set of photos. As the first step, emotion categories are automatically extracted from the input music and photos. Based on Hevner's work [6], our emotion taxonomy consists of eight emotion classes: *sublime*, *sad*, *touching*, *easy*,

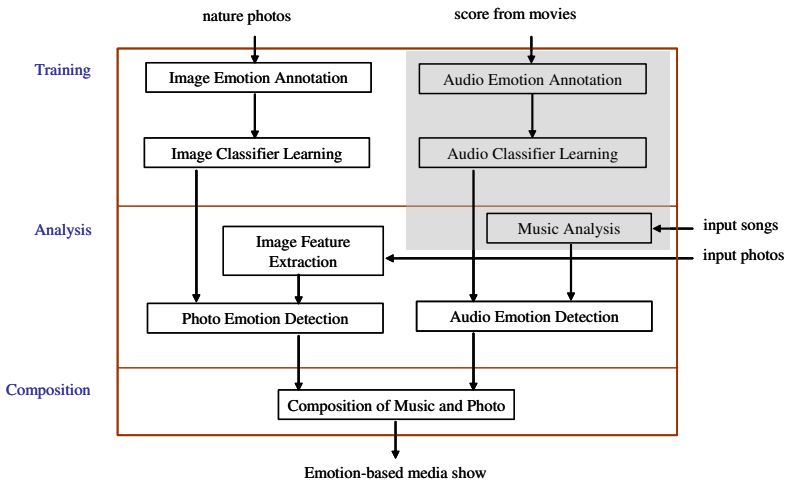

Fig. 1. System framework for emotion-based media show.

light, *happy*, *exciting* and *grand*. For more details on the taxonomy, please refer to Wu and Jeng's paper [2]. For automatic music emotion detection, we directly apply Wu and Jeng's method [2]. In their method, the music features are selected by F-scores and a multi-class classifier is built by SVM. The average precision and recall are around 70%.

For automatic photo emotion detection, at the training stage, we collected and annotated a set of natural photos. From these images and their annotations, the distribution of emotion class is computed and classification models are learned. At the analysis stage, input photos are analyzed to predict their emotion classes by using the classifiers and models learned at the training stage. At the composition stage, based on beat analysis and music emotion detection, music is first divided into several segments, each of which contains a homogeneous emotional expression. We then attempt to assign a photo to each segment so that both have the same emotion expression. In addition to high-level emotional labels, two more criteria are considered to increase the harmony between music and photos. The first criterion is the harmonization between timbre of the music and color of the images. The second criterion is the temporal visual coherence of the photo sequence. Based on these criteria, photo sequence selection is formulated as an optimization problem and solved by a greedy algorithm.

3 Photo Emotion Detection

For automatic photo emotion detection, we collected and annotated a set of natural photos (Section 3.1). From them, Bayesian classifiers are learned. For a photo to be categorized, three types of visual features are extracted (Section 3.2) and fed into classifiers to determine its emotion category (Section 3.3).

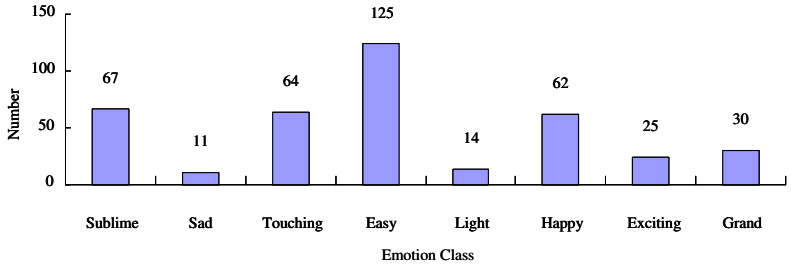

Fig. 2. The emotion class statistics of the image database.

3.1 Training data collection

There are few image datasets with emotion annotations. In psychology, Lang conducted a series of experiments [7] using a database of photographs, the International Affective Picture System (IAPS), as the emotional stimulus. Unfortunately, IAPS has a different emotion taxonomy. In addition, most photos in IAPS are not what we will experience in our daily life. Thus, we collected 398 photos from internet and had them subjectively labeled by 496 participants through an on-line tagging system. To make the presentation joyful to watch and avoid dealing with high-level semantics, we selected photos according to two criteria: (1) we prefer images related to our daily life and without specific semantic meaning; and (2) we avoid photos containing human faces because facial expressions likely dominate the moods of photos. As a result, most photos are scenic photographs containing forrest, mountains, beaches, buildings and so on. There are also some photos on activities such as surfing and parties.

To be consistent with music emotion detection, we used the same tagging interface of emotion annotation as Wu and Jeng's work [8] for labeling photos. Because each photo is labeled by many users and they could perceive different emotions to it, instead of a single label, we represent the aggregated annotation of an image as a distribution vector over the eight emotion classes. To show some statistics about the dataset, Fig. 2 gives the numbers of images for all eight emotion classes by assigning each image to the dominant class in its distribution. Fig. 3 shows representative photos from each emotion class.

3.2 Visual features

To construct a good computational model for photo emotion detection, we need to select a set of visual features which effectively reflect emotions. Except for high-level semantics which is still difficult to annotate accurately, among low-level visual features related to emotion, color is probably the most important one. Generally speaking, warm colors bring viewers warmth, excitement of emotions or even anger, while images dominated by cool colors tend to create cool, clamming, and gloomy moods. In some situations, a great deal of detail gives a

Fig. 3. Sample images with different emotion tags from collection.

sense of reality to a scene, and less detail implies more smoothing moods. As a result, textureness of an image also affects the viewer's emotion. Moreover, the directions of lines can express different feelings. Strong vertical elements usually indicate high tensional states while horizontal ones are much more peaceful. In this paper, we characterize three visual features for photo emotion detection: color, textureness, and line.

Color. The HSV color space is more suitable for human perception. We apply the method of Zhang *et al.* [9] to quantize the color space into 36 non-uniform bins. H channel is considered differently from S and V channels in a more similar way to the human vision model.

Textureness. Texture is an important cue for image feelings. In this paper, as a measure for local textureness, an entropy value is calculated for each pixel from the CIE Lab color histogram of its 32×32 neighborhood. Then, the image is divided 4×4 blocks and an average entropy value is calculated for each block.

Line. To model the line properties in images, edge features are obtained by a Canny edge detector, and then a wavelet transform is applied. The energy of spatial graininess, vertical stripes and horizontal elements are evaluated from moments of wavelet coefficients at various frequency bands.

3.3 Photo emotion detection

Because the annotated images are too few to cover the visual feature space, we can't obtain good classification performance by directly applying supervised learning approaches such as SVM. Instead, we propose a Bayesian approach to use unlabeled data to boost classification performance. We randomly selected 2,000 unlabeled images from another database, IAPR [10], of natural images. Visual features are extracted from these 2,000 images. Affinity propagation al-

gorithm [11], a recently introduced powerful unsupervised clustering method, is used to group similar features into the same cluster. Assume that m clusters are formed and X_i is the representative vector of the i^{th} cluster. From the clustering results, we adopt the package of support vector machine (SVM), LIBSVM [12], to construct a model M_i by feeding all features in the i^{th} cluster and fivefold number of features in other clusters. The SVM model is used to predict the probability $p(X_i|I, M_i)$ that a new image I belongs to the i^{th} cluster given its visual feature. We applied a standard sigmoid function to convert the output margin of SVM to a probability.

As explained earlier, emotion annotation for an image is expressed as a distribution over the eight defined emotions. For these 398 annotated photos, we use affinity propagation again to cluster the annotations into several discriminating clusters according to their emotional distributions. The clustering results represent major emotional distribution types. In our implementation, 19 clusters are generated, implying there are 19 major emotional distribution types. E_j denotes the representative distribution of the j^{th} major emotional distribution type.

At the analysis stage, given a novel image I_t, we use the following formula to classify I_t into one of the 19 major emotional distributions.

$$E_t = \arg\max_{E_j} p(E_j|I_t) = \arg\max_{E_j} \sum_{i}^{m} p(E_j|X_i)p(X_i|I_t, M_i), \tag{1}$$

where $p(X_i|I_t, M_i)$ is the probability that I_t belongs to the feature cluster X_i, estimated by the model M_i, and $p(E_j|X_i)$ is prior indicating the probability of belonging to the major emotional distribution E_j given that an image belongs to the feature cluster X_i. To evaluate $p(E_j|X_i)$, we adopt Bayes' rule

$$p(E_j|X_i) \propto p(X_i|E_j)p(E_j),$$

where $p(E_j)$ is the prior probability for the major emotional distribution E_j (Fig. 2), and $p(X_i|E_j)$ indicates the probability that an image's visual feature belongs to cluster X_i given that the image belongs to the major emotional distribution E_j. Assuming that P_{E_j} denotes the set of photos belonging to the major emotional distribution E_j in the training set, $p(X_i|E_j)$ is then approximated as

$$p(X_i|E_j) = \frac{1}{|P_{E_j}|} \sum_{I \in P_{E_j}} p(X_i|I, M_i).$$

Finally, we assign the test image I_t to the emotion category having the maximal probability in the emotional distribution E_t. By using this Bayesian framework, we can use help from unlabeled data and work around the problem of not having sufficient photos with emotion labels. From preliminary experiments, the accuracy of Bayesian framework is 43%. Although the number does not look impressive, mis-classified photos are often classified as nearby emotions. Therefore, it is unlikely to assign an opposite emotion to a photo. Thus, even if not extremely accurate, for our emotion-based music visualization, the detection results are good enough because music and photos are similar in emotion even if they could be labeled differently.

4 Composition of Music and Photos

In this section, we describe the method for music visualization composition. The music is first divided into several segments of homogenous emotional expressions. Next, for each music segment, we have to select a photo with the same emotion. In order to express various emotions in music, the number of photos is preferred to be large enough to cover a variety of emotions. As discussed in the previous section, the input photos are first grouped into different emotion classes. For each music segment, we must select a photo whose emotion label is the same as the music segment. Since we have many photos in an emotion category, we have the luxury to select a better one from multiple photos with the same emotion. Thus, in addition to matching high-level emotion category, we consider two additional criteria to further improve coordination between music and photos: (1) the consistence between low-level features of music and photos, and (2) visual coherence between successive photos.

4.1 Music segmentation

A song often consists of more than one emotion. Thus, to detect emotion, we have to divide the song into several independent sections, each of which contains a homogeneous emotion. However, it is difficult to define where emotions switch. In fact, different emotion transition boundary decision strategies may result in different emotion detection results. Here, we present a three-stage analysis method to divide music into segments.

First, beat tracking is used to detect the rhythmic pulse of a piece of music. It is the salient part of music we often notice. As a result, the beat time is the good time to switch photos because synchronization of photo switching to music beats often improves enjoyment of watching the visualization. We use the beat tracking algorithm by Dixon [13] to divide music into many clips. Because the durations between beats are often too short for music emotion detection, we group adjacent clips into five-second segments according to the beat information. The boundary of each segment is picked to be as close to the beat time as possible. In our system, segments are the basic units from which the emotions are extracted and each segment will be associated with a photo in the visualization. Thus, the music emotion detection method is used to extract an emotion label for each segment.

After we obtain the emotion label of each segment, a regrouping process is invoked to find better emotion transition boundaries. Successive segments with the same emotion are grouped together as a longer section (Fig. 4). The emotion transition boundaries between sections are then further refined with the method proposed by Lu *et al.* [1].

4.2 Photo sequence selection

For music visualization composition, after finding segments of the music, we have to assign a photo to each segment. The input photos are first classified into different emotion categories. For a section S with an emotion label e, assume that S

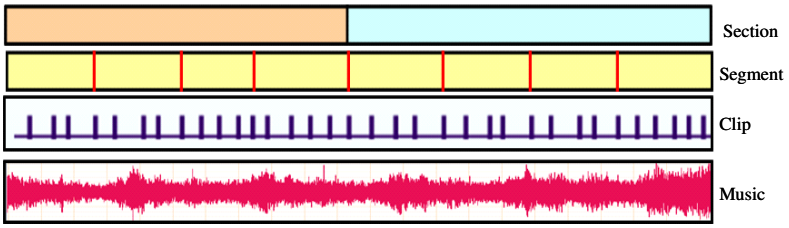

Fig. 4. The music segmentation process.

has n segments and is denoted as $S = \langle s_1, s_2, \ldots, s_n \rangle$ where s_i is the i^{th} segment of S. For the emotion class e, we have a set of m photos, $P_e = \{p_1, p_2, \ldots, p_m\}$, which has the same emotion label e. The photo sequence selection problem is to find a photo sequence $o = \langle \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n \rangle$ for section S, where $\hat{p}_i \in P_e$ and photo $\hat{p}_i$ is displayed when music segment s_i is played. The photo sequence selection is performed for each section independently. To select a proper sequence, in addition to matching high-level emotion notions, we further include the following two criteria to enhance the harmony between music and photo sequence.

Criterion 1: the coordination between low-level features of the music and photos. To be more consistent in visual and auditory perception, some low-level attributes of music and photos are selected to coordinate each other. As we listen to music, its timbre greatly affects our feelings to it. Qualities of timbre are often described as an analogy to color, since timbre is perceived and understood as a gestalt impression reflective of the entire sound. Ox pointed out that timbre is an essential feature to music in a similar way that color is to a painting [14]. Indeed, timbre literally means the color of sound. Hence, timbre is selected as the music attribute to be matched with the color of photos. Among the timbrel features, spectral centroid and spectral flux are important perceptual properties in the characterization of musical timbre [15]. According to these clues, we define a fitness function D_{feature} for measuring the coordination between the timbre of a music segment s_i and the color of a photo p as

$$D_{\text{feature}}(s_i, p) = \omega_{\text{centroid}} \cdot d_{\text{centroid}}(s_i, p) + \omega_{\text{flux}} \cdot d_{\text{flux}}(s_i, p) \ .$$

The function d_{centroid} measures the distance between spectral centroid of a music segment and brightness of a photo. If the centroid of music is large, we prefer a brighter photo. And the function d_{flux} expresses the distance between spectral flux of music and color contrast of a photo. If the flux variation in one segment is large, it means that the variation of the music timbre is large. Thus, a photo with higher color contrast is more suitable to be displayed with the segment.

Criterion 2: the temporal visual coherence of the photo sequence. We prefer a sequence whose visual appearance changes gradually along time. To measure temporal coherence of a photo sequence, the Lab color space is used. A histogram

is computed for each photo in the a,b space, and we measure the distance between a pair of photos as the chi-squared distance between their ab-histograms.

$$D_{\text{coherence}}(p_i, p_j) = \begin{cases} d_{\chi^2}(H_{ab}(p_i), H_{ab}(p_j)) & \text{if } p_i \neq p_j \\ \infty & \text{otherwise} \end{cases}$$

where $H_{ab}(\cdot)$ is the histogram constructed in a,b space and $d_{\chi^2}(\cdot)$ is the chi-squared distance. An infinity distance is assigned if p_i and p_j are the same photo. This is to prevent the case that the same photo appears in succession.

The problem of photo sequence selection can then be modeled as selecting an optimal sequence $o = \langle \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n \rangle$ that minimizes the following function,

$$D(o) = \alpha \sum_{i=1}^{n} D_{\text{feature}}(s_i, \hat{p}_i) + \beta \sum_{i=1}^{n-1} D_{\text{coherence}}(\hat{p}_i, \hat{p}_{i+1}), \tag{2}$$

where α and β are weighting parameters, D_{feature} and $D_{\text{coherence}}$ are defined above to measure the harmony between music and photos and the temporal visual coherence of a photo sequence.

It is time-consuming to solve the optimization by exhaustive search. Instead, a greedy algorithm is used to find a local optimum solution. Let the mapping $\Phi(i) = j$ denote that the optimal sequence selects photo p_j for segment s_i. The greedy algorithm sequentially finds the optimal mapping from the first segment to the last segment. Assuming that we already have solved up to the i^{th} segment. That is, the values of $\Phi(1), \cdots, \Phi(i)$ are solved. Then, we evaluate $\Phi(i+1)$ as

$$\Phi(i+1) = \arg\min_j \alpha D_{\text{feature}}(s_{i+1}, p_j) + \beta D_{\text{coherence}}(p_{\Phi(i)}, p_j).$$

Locally best photos are sequentially found from the first segment to the last segment by repeatedly evaluating the above equation. Finally, the sequence $o = \langle p_{\Phi(1)}, p_{\Phi(2)}, \ldots, p_{\Phi(n)} \rangle$ is returned as the photo sequence to be played with the music section. All sections are processed in the same way independently to generate the whole visualization for the input music.

5 Evaluation Results

Since emotional perception of acoustic and visual media is subjective, objective evaluation of our system is difficult. Thus, we evaluate our system through a subjective user evaluation. Three types of music visualization are compared. In addition to the one generated by our system, we generate another visualization by randomly selecting photos without considering emotions. The third one is the visualization from Microsoft media player. Twenty-one evaluators aged from 20 to 50 were invited to evaluate the presentations. Particularly, we focus on evaluating how well our system bridges and enhances perception of music and photos through coordinating their emotional expressions. The following questions were asked.

Question 1 coordination: How do you think of the connection between music

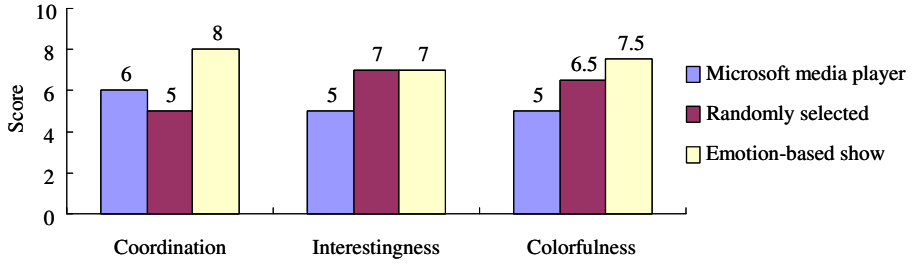

Fig. 5. Results of user evaluation.

and photos?
Question 2 interestingness: How do you think of the presentation style?
Question 3 colorfulness: How well do you think the visual contents enrich the audio?
Evaluators were asked to give scores from 1 to 10 (higher scores for more contentment) to show their satisfaction with these presentations. Fig. 5 summarizes the evaluation results. The presentation[†] generated by our method consistently has the highest scores for all questions. From the evaluation, we find that music visualization using photos is more interesting to watch than the animated patterns in traditional music visualization. Even the presentation with randomly selected photos is more fun to watch than media player. Our emotion-based music visualization coordinates the presence of music and photos much better than media player as our visualization captures higher-level notions of media in addition to low-level features. Thus, the music listening experience is enriched by synchronized emotional expressions of photos.

6 Conclusion and Future Work

In this paper, we have proposed a framework for creating emotion-based music visualization. To achieve this goal, we have collected and annotated 398 photos and proposed an automatic photo emotion detection method based on a Bayesian framework. For composition of music and photos, in addition to high-level emotion notions, we also consider temporal coherence of photo sequence and coordinate the presence of music and photos by their low-lever features of timbre and color. User evaluations indicates that emotion-based music visualization effectively enriches music listening experience.

There are several research avenues that we would like to explore in the future. For example, there is certainly room for improvement on the accuracy of emotion detection algorithms. Some transition effects could be added to harmonize with emotions. With little modification, our method could be used to create emotion-based photo slideshows. In addition to matching emotion categories,

[†] It can be found at `http://www.csie.ntu.edu.tw/~b90030/emotionbased.wmv`.

the composition method described in Section 4 could also be applied to other problem domains, such as situation types. Furthermore, the proposed method could be embedded into digital photo displays, making them more fun to watch.

Acknowledgment

This paper is primarily based on the work supported by the National Science Council (NSC) of Taiwan, R.O.C., under contracts NSC95-2752-E-002-006-PAE and NSC95-2622-E-002-018.

References

1. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Transactions on Audio, Speech & Language Processing **14**(1) (2006) 5–18
2. Wu, T.L., Jeng, S.K.: Probabilistic estimation of a novel music emotion model. In: The 14th International Multimedia Modeling Conference, Kyoto, Japan (2008)
3. Chen, J.C., Chu, W.T., Kuo, J.H., Weng, C.Y., Wu, J.L.: Tiling slideshow. In: ACM Multimedia, Santa Barbara, CA, USA (2006)
4. Hua, X.S., Lu, L., Zhang, H.J.: Automatic music video generation based on temporal pattern analysis. In: ACM Multimedia, New York, NY, USA (2004)
5. Mulhem, P., Kankanhalli, M.S., Yi, J., Hassan, H.: Pivot vector space approach for audio-video mixing. IEEE MultiMedia **10**(2) (2003) 28–40
6. Hevner, K.: Expression in music: a discussion of experimental studies and theories. Psychol. Rev. **42** (1935) 186–204
7. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International affective picture system (iaps): technical manual and affective ratings. NIMH Center for the Study of Emotion and Attention (1997)
8. Wu, T.L., Jeng, S.K.: Regrouping expressive terms for musical qualia. In: WOCMAT on Computer Music and Audio Technology, Taiwan (2007)
9. Zhang, L., Lin, F., Zhang, B.: A cbir method based on color-spatial feature. In: IEEE Region 10 Annual International Conference. (1999)
10. Grubinger, M., Clough, P., Muller, H., Deselears, T.: The iapr tc-12 benchmark – a new evaluation resource for visual information systems. In: International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval. (2006)
11. Frey, B.J.J., Dueck, D.: Clustering by passing messages between data points. Science (January 2007)
12. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
13. Dixon, S.: Mirex 2006 audio beat tracking evaluation: Beatroot. MIREX (2006)
14. Ox, J.: Two performances in the 21st century virtual color organ: Gridjam and im januar am nil. In: Proceedings of the Seventh International Conference on Virtual Systems and Multimedia. (2001) 580
15. Grey, J.: An exploration of musical timbre. Ph.D. Dissertation, Stanford University (1975)