Analyzing Emotional Semantics of Abstract Art Using Low-Level Image Features

He Zhang[1], Eimontas Augilius[1], Timo Honkela[1], Jorma Laaksonen[1], Hannes Gamper[2], and Henok Alene[1]

[1] Department of Information and Computer Science
[2] Department of Media Technology
Aalto University School of Science, Espoo, Finland
{he.zhang,eimontas.augilius,timo.honkela,jorma.laaksonen, hannes.gamper,henok.alene}@aalto.fi

Abstract. In this work, we study people's emotions evoked by viewing abstract art images based on traditional low-level image features within a binary classification framework. Abstract art is used here instead of artistic or photographic images because those contain contextual information that influences the emotional assessment in a highly individual manner. Whether an image of a cat or a mountain elicits a negative or positive response is subjective. After discussing challenges concerning image emotional semantics research, we empirically demonstrate that the emotions triggered by viewing abstract art images can be predicted with reasonable accuracy by machine using a variety of low-level image descriptors such as color, shape, and texture. The abstract art dataset that we created for this work has been made downloadable to the public.

Keywords: emotional semantics, abstract art, psychophysical evaluation, image features, classification.

1 Introduction

Analyzing image emotional semantics is an emerging and promising research direction for Content-Based Image Retrieval (CBIR) in recent years [5]. While the CBIR systems are conventionally designed for recognizing object and scene such as plants, animals, people etc., an Emotional Semantic Image Retrieval (ESIR) [17] system aims at incorporating the emotional reflections to enable queries like "beautiful flowers", "lovely dogs", "happy faces" etc. In analogy to the concept of *semantic gap* implying the limitations of image recognition, the *emotional gap* can be defined as "the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal" [6].

Though emotions can be affected by various factors like gender, age, culture, background, etc. and are thus considered as high-level cognitive processes, they still have certain stability and generality across different people and cultures [13]. This enables researchers to generalize their proposed methodologies from limited

J. Gama, E. Bradley, and J. Hollmén (Eds.): IDA 2011, LNCS 7014, pp. 413–423, 2011.

samples given a sufficiently large number of observers [17]. Due to the subjectivity of emotions, two main challenges in analyzing image emotional semantics can be identified: 1) measuring effectively the subjectivity or consensus among people for obtaining robust ground-truth labels of the shown image. 2) extracting informative image content descriptors that can reflect people's affective feelings evoked by the shown image.

As for measuring the subjectivity, adjective (impression) words are often selected to represent people's emotions at first. Recently, many researchers use opposing adjective word pairs to represent emotional semantics, such as happy-sad, warm-cold, like-dislike etc. For example, respectively 9, 10, and 12 adjective word pairs were used in [19], [13], and [18] for constructing the emotional factor space, with a number of 12, 31, and 43 observers involved per each. Generally, using a large number of adjective word pairs may improve the experimental results. However, this will increase the evaluation time and affect the observers' moods, which in turn would lower the generality of evaluation results. Besides, the adjectives often conceptually overlap in adjective space [8].

As for extracting meaningful image descriptors, many attempts have been reported (e.g. [4,9,16,18,19]). Most of the works developed features that are specific to the domains related to art and color theories, which lacks generality and makes it difficult for researchers who are unfamiliar with computer vision algorithms to perform image analysis on their own data [15]. Besides, the image datasets used in these works are mainly scenery and photographic images containing recognizable objects and faces which are likely to distort people's emotions.

In this article, we study people' emotional feelings evoked by viewing abstract art images within a classification framework. Similar to [9], we choose abstract art images as our dataset since they usually contain little contextual information that can distort the observer's emotions. However, the features used in [9] are designed on art and color theories and thus are deficient in generality. In contrast, we seek to bridge the *emotional gap* by utilizing a large variety of conventional low-level image features extracted from raw pixels and compound image transforms, which have been widely used in computer vision and machine learning problems. Compared with a designed baseline method, we can achieve significantly better results on classifying image emotions between *exciting* and *boring*, *relaxing* and *irritating*, with a relatively small number of image samples. The features used in this article have been implemented by Shamir et al [15] as an open source software and the abstract art image dataset that we collected has been made downloadable to the research community [1].

Section 2 describes the collected dataset and online psychophysical evaluations. Section 3 introduces the low-level image features extracted from both the raw pixels and image compound transforms. The classification setups are explained in Section 4. Section 5 presents the experimental results with analysis. Finally the conclusions and future work are made in Section 6.

[1] http://research.ics.tkk.fi/cog/data/esaa/

2 Data Acquisition

Data collection is important in that it provides art images for evaluation by real observers such that both the input samples and their ground-truth labels can be obtained for training and testing the classifier in the latter stage. The duration of evaluation process should not be too long, otherwise the observers will get tired and their responses will deteriorate. This means we should select a limited number of image samples, without affecting as much the generality of the model as possible.

2.1 Image Collections

We collected 100 images of abstract art paintings with different sizes and qualities through Google image search. These abstract art paintings were created by artists with various origins and the image sizes were within a range between 185×275 and $1,000 \times 1,000$. We kept the image samples the same as they were initially selected from Internet and no data preprocessing such as image downsampling or cropping were performed. This not only mimics a real user's web browsing scenario, but also it facilitates the image descriptors in extracting discriminative information from the art images in the latter stage.

2.2 Psychophysical Evaluations

Semantic Differential (SD) [12] is a general method for measuring the affective meaning of concepts. Each observer is asked to put a rating on a scale between two bipolar adjective words (e.g. happy-sad) to describe his or her emotions evoked by the shown images. The final rating of every opposing adjective word pairs for an image is obtained by averaging ratings over all the observers.

Following the SD approach, an online image survey was conducted through our designed web user interface[2]. To lower the bias, 10 female and 10 male observers both Asians and Europeans were recruited. All the observers have a university degree and their ages are between 20 and 30 years old. During the evaluation, each observer was shown 100 abstract art images with one per page. Under each image, he or she was asked to indicate ratings for both *exciting-boring* and *relaxing-irritating*. For every word pair, we used 5 ratings with respective values of $-2, -1, 0, 1, 2$ to denote an observer's affective intensity. Thus the overall rating of each adjective pair for an image was the average rating score from all the observers. The evaluation took about 10 to 15 minutes in general, which was easily acceptable for most of the participants.

Post-processing: To obtain the ground-truth labels, we adopted a simple rule: if an image, for an adjective word pair, received an average rating score larger than zero, then it was treated as a positive sample for the classifier; if the average score was smaller than zero, it is a negative sample. Figure 1 shows the top 3 images for each impression word, after sorting the 100 images by their average

[2] `http://www.multimodwellbeing.appspot.com/?controlled`

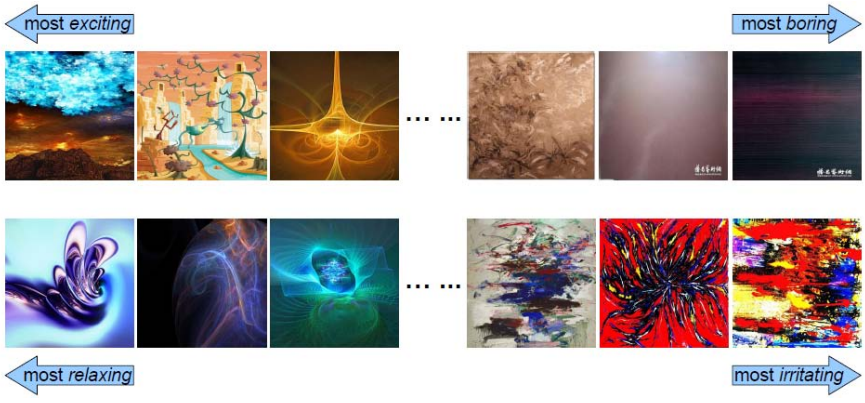

Fig. 1. The top 3 images sorted by the average scores over 20 observers for *exciting-boring* (upper row) and *relaxing-irritating* (bottom row)

ratings over 20 observers for the 2 adjective word pairs respectively. Although there still lacks a common measurement of subjectivity, intuitively our survey results revealed certain generalities towards abstract art paintings among people.

3 Feature Extraction

For describing the visual art paintings, a large set of image features have been extracted from both the raw image and several compound image transforms, which were found highly effective earlier in biological image classification and face recognition [11], as well as recognition of painters and schools of art [14]. The (eleven) groups of features are listed in Table 1. In addition to the raw pixels, the image features have also been extracted from several transforms of the image and transforms of transforms [14]. These transforms are Fast Fourier Transform

Table 1. The features used in [14] and our study

Group of Features	Type	Dimension
First four moments	Image Statistics	48
Haralick features	Texture	28
Multiscale histograms	Texture	24
Tamura features	Texture	6
Radon transform features	Texture	12
Chebyshev statistic features	Polynomial	400
Chebyshev-Fourier features	Polynomial	32
Zernike features	Shape & Edge	72
Edge statistics features	Shape & Edge	28
Object statistics	Shape & Edge	34
Gabor filters	Shape & Edge	7

(FFT), Wavelet 2D Decomposition, Chebyshev Transform, Color Transform, and Edge Transform, as shown in Figure 2.

The idea in the transforms mentioned above is to create automatically a numerical representation that captures different kinds of basic qualities of the images in a reasonably condensed representation. The dimensionality of a $1,000 \times 1,000$ RGB color image can be reduced from $3,000,000$ to about $4,000$ with an increased descriptiveness of the features (see below for details). For instance, transforms like FFT or Wavelets can detect invariance at different levels of detail and help in reducing noise to benefit the classification.

The total number of the numeric image descriptors used in this article is 4,024 for every image, whereas the authors in [14] excluded the *Zernike features*, *Chebyshev statistics*, and *Chebyshev-Fourier features* from several compound transforms, resulting in a total number of 3,658 descriptors.

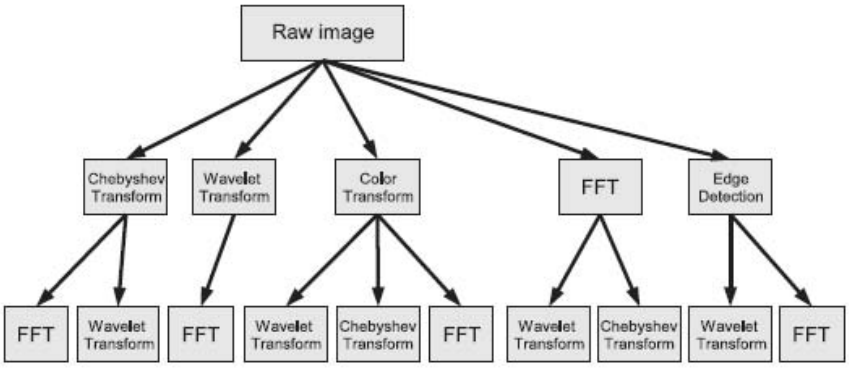

Fig. 2. Image transforms and paths of the compound transforms described in [14]

The image features and transforms above have been implemented as part of an open source software [15] and we directly utilize it in our study. For details of the feature extraction algorithms, one may refer to [10] and the references therein.

4 Classification of Emotional Responses

After calculating all the features for all images, a mapping needs to be built to bridge the semantic gap between the low-level image features and the high-level emotional semantics.

Feature Selection: Due to the usage of the large number of image descriptors, a feature selection step is expected prior to the recognition stage, since the discriminative power may vary in features and only some of them are informative, whereas others are redundant and/or non-related. Various feature selection

strategies exist. Here we utilize the popular Fisher score [1] for feature ranking, i.e. assigning each feature f with a weight W_f such that:

$$W_f = \frac{\sum_{c=1}^{C}(\bar{f} - \bar{f}_c)}{\sum_{c=1}^{C}\sigma_{f,c}^2}, \tag{1}$$

where C is the total number of classes ($C = 2$ in our case); $\bar{f}$ denotes the mean of feature f in the whole training set; $\bar{f}_c$ and $\sigma_{f,c}^2$ denote respectively the mean and variance of feature f among all the training images of class c; The Fisher score can be explained as the ratio of between-class variance to within-class variance. The larger the F-score is, the more likely that this feature is more discriminative. Therefore the features with higher Fisher scores are retained whereas those with lower values are removed.

Classification: Support Vector Machine (SVM) [3] is chosen to build the mapping as it is the state-of-art machine learning method and has been used for classification in recent emotion-related studies [19,4]. In our paper, we use the SVM package LIBSVM [2] with default parameters in order to ensure reproducibility of the experimental results. After splitting the image dataset into training and testing sets, an SVM classifier is learned using the features of training set (consisting of both positive and negative image samples). Then for every image in the testing set, a corresponding class label or affective category is automatically predicted.

Evaluation: To measure the performance of SVM, the classification *accuracy* is calculated, defined as the proportion of correctly predicted labels (both positive and negative) within testing image set. Another measure is the *precision* or positive predictive rate, since we are more interested in how many positive image samples can be correctly recognized through a machine learning approach.

5 Results and Analysis

The LIBSVM [2] package was utilized in all the experiments, with linear kernel and default parameters. Since a "standard" baseline method does not exists in this field, we generated for each image an array of 4024 *random* numeric values, and repeated the same training and testing procedure as described in Section 4. This facilitates us to validate the effectiveness of representing emotional semantics with *real* low-level image features. For all the binary classification cases, the number of positive image samples was roughly equal to that of the negative ones. In each case, we calculated the final classification accuracies and precisions based on 5-fold cross validation.

Classification Performances: Figure 3 shows the average precisions and accuracies as a function of the percentages of best real image features, compared with those using random image features in all the 6 cases. Table 2 lists for each case the best average precisions and accuracies with the corresponding percentages of the best image features, compared with the respective precisions and accuracies

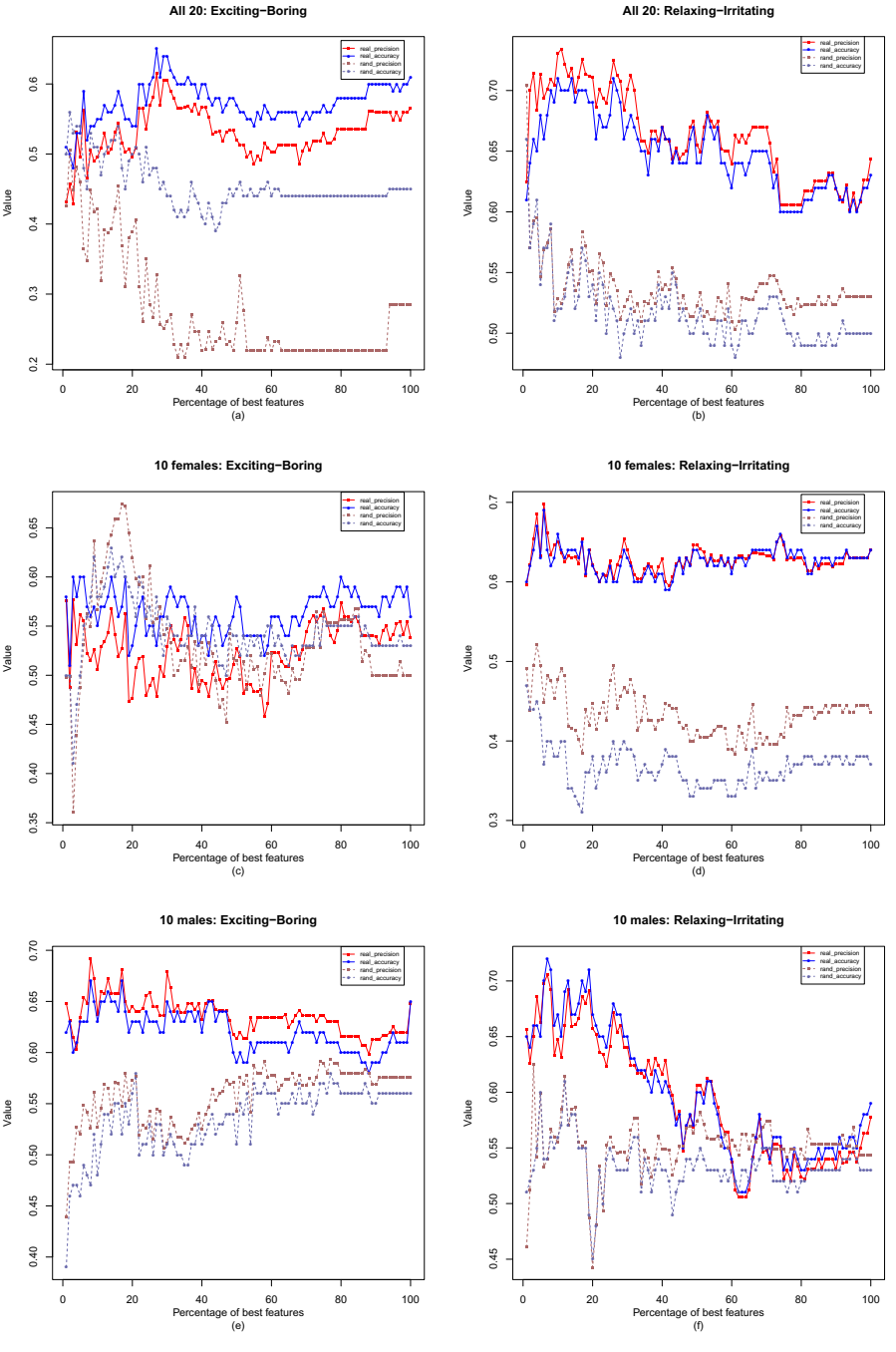

Fig. 3. The average precisions and accuracies as a function of the percentage of best real image features, compared with those using random features in all 6 cases

Table 2. The best average precisions (column 3) and accuracies (column 5) with the corresponding percentages (column 2) of the best image features (denoted as **Real**), compared with the respective precisions (column 4) and accuracies (column 6) at the same percentages of random image features (denoted as **Rand**) in the 6 cases (column 1). All statistics are shown in percentages (%).

Case	Best	Pre.-Real	Pre.-Rand	Acc.-Real	Acc.-Rand
All 20: Exciting-Boring	27	**62**	33	**65**	48
All 20: Relaxing-Irritating	11	**76**	53	**70**	52
10 female: Exciting-Boring	3	**57**	49	**58**	50
10 female: Relaxing-Irritat.	6	**70**	48	**69**	40
10 male: Exciting-Boring	8	**69**	56	**67**	47
10 male: Relaxing-Irritat.	7	**72**	53	**71**	54
Average	10	**68**	49	**67**	49

at the same percentages of random image features. Generally, the classification performances of using real image features significantly outperform those of using the random ones, except for the case in Figure 3(c). The average precision using real image features over all the 6 cases was 19% higher than that using the random features, using an average 10% of the best real image features. For the case in Figure 3(b), a peak precision (accuracy) of 0.76 (0.70) was obtained when using the best 11% of real image features sorted by their Fisher scores, compared with a precision (accuracy) of 0.53 (0.52) at the same percentage when using the random ones. Similar comparisons can be made in the other 5 cases.

Rankings of Feature Groups: Table 3 lists the top 10 feature groups for the 2 cases in "All 20" (within the best 10% of real image features).

For the "Exciting-Boring" case, the best 2 feature groups are Color histogram extracted from raw images and Multiscale histogram from Chebyshev FFT compound transforms, whereas for the "Relaxing-Irritating" case, the best 2 feature groups are Edge statistics extracted from raw images and Radon transform from Color FFT compound transforms. This conforms to the previous research using

Table 3. The top 10 feature groups (compound transforms) in "all 20" cases

Exciting-Boring	Relaxing-Irritating
Color histograms	Edge statistics
Multiscale histograms (Chebyshev FFT)	Radon transform (Color FFT)
Radon transform	Haralick texture (Wavelet)
Chebyshev statistics (Color Chebyshev)	Tamura texture (FFT Wavelet)
Haralick texture (Color)	Haralick texture (Chebyshev FFT)
Tamura texture (Edge)	Tamura texture (Color FFT)
Chebyshev-Fourier (Color)	Tamura texture (Wavelet FFT)
Chebyshev statistics (Color)	Tamura texture (Color)
Tamura texture (Edge Wavelet)	Tamura texture (FFT)
Chebyshev-Fourier (FFT Wavelet)	Radon transform (Edge)

features based on color theories for image emotions' classification [9,16], and to the study where edge and texture features were favored for different art styles' recognition [14].

6 Conclusions and Future Work

In this work, we have studied people's emotions evoked by viewing abstract art images within a machine learning framework. The ground-truth labels of sample images were obtained by conducting an online web survey, where 20 observers both females and males were involved in evaluating the abstract art dataset. A large variety of low-level image descriptors were utilized to analyze their relationship with people's emotional responses to the images. Both the utility implementing the image features and the abstract art images that we collected for the experiments are in public domain.

Our results show that the low-level image features, instead of domain specific ones, can be used for effectively describing people's high-level cognitive processes, which has been empirically demonstrated in our experiments that the image emotions can be well recognized in classification tasks. Besides, by examining the rankings of feature groups sorted by their Fisher scores, the most discriminative features are color, shape, and texture, which in itself conforms to the art and color theories, as well as to several recent studies related to image emotional semantics (e.g. [14]). Even in the case of abstract art images where the semantic content of the image does not influence the evaluation, a high degree of subjectivity is involved. This is actually true even for linguistic expressions [7]. Therefore, the aim is not to create a "correct" classification model for the emotional responses but rather model the tendencies in the evaluation. When there are thousands of subjects involved in this kind of study, it is possible to model in additional detail the agreements and disagreements in the subjective evaluations and potentially associate those with some variables that characterize each individual.[3]

Our next step is to compare the low-level features with the domain specific ones. A direct application is to integrate the low-level image features into an ESIR system to facilitate the emotional queries. Still, more advanced feature selection algorithms can be tested since the Fisher's criterion neglects the relationships between features. Besides, other relevance feedback modalities [20] from the observer, such as eye movements and speech signals (see [21] for instance), could be combined with image features to enhance the recognition performance, so that a deeper understanding of image emotions can be accomplished. In addition to Emotional Semantic Image Retrieval [17], potential applications include automatic selection of images that could be used to induce specific emotional responses.

Acknowledgements. This work is supported by Information and Computer Science Department at Aalto University School of Science. We gratefully

[3] For this purpose, you are welcome to interact with our online survey at `http://www.multimodwellbeing.appspot.com/?controlled`

acknowledge Ms. Na Li and her colleagues for attending the psychophysical evaluations. We are grateful for the EIT ICT Labs and Dr. Krista Lagus for the fact the Wellbeing Innovation Camp 2010 served as the starting point for the work reported in this paper. We wish that the increasing understanding of the wellbeing effects of pieces of art and other cultural artefacts will approve to have useful applications in the future.

References

1. Bishop, C.M.: Pattern recognition and machine learning, vol. 4. Springer, New York (2006)
2. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
4. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part III. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR) 40(2), 5 (2008)
6. Hanjalic, A.: Extracting moods from pictures and sounds: Towards truly personalized TV. IEEE Signal Processing Magazine 23(2), 90–100 (2006)
7. Honkela, T., Janasik, N., Lagus, K., Lindh-Knuutila, T., Pantzar, M., Raitio, J.: GICA: Grounded intersubjective concept analysis – a method for enhancing mutual understanding and participation. Tech. Rep. TKK-ICS-R41, AALTO-ICS, ESPOO (December 2010)
8. Honkela, T., Lindh-Knuutila, T., Lagus, K.: Measuring adjective spaces. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010, Part I. LNCS, vol. 6352, pp. 351–355. Springer, Heidelberg (2010)
9. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: Proceedings of the International Conference on Multimedia, pp. 83–92. ACM, New York (2010)
10. Orlov, N., Johnston, J., Macura, T., Shamir, L., Goldberg, I.G.: Computer Vision for Microscopy Applications. In: Obinata, G., Dutta, A. (eds.) Vision Systems–Segmentation and Pattern Recognition, pp. 221–242. ARS Pub. (2007)
11. Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D.M., Goldberg, I.G.: WND-CHARM: Multi-purpose image classification using compound image transforms. Pattern Recognition Letters 29(11), 1684–1693 (2008)
12. Osgood, C.E., Suci, G.J., Tannenbaum, P.: The measurement of meaning. University of Illinois Press, Urbana (1957)
13. Ou, L., Luo, M.R., Woodcock, A., Wright, A.: A study of colour emotion and colour preference. Part I: Colour emotions for single colours. Color Research & Application 29(3), 232–240 (2004)
14. Shamir, L., Macura, T., Orlov, N., Eckley, D.M., Goldberg, I.G.: Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. ACM Transactions on Applied Perception 7, 8:1–8:17 (2010)
15. Shamir, L., Orlov, N., Eckley, D.M., Macura, T., Johnston, J., Goldberg, I.G.: Wndchrm–an open source utility for biological image analysis. Source Code for Biology and Medicine 3(1), 1–13 (2008)

16. Solli, M., Lenz, R.: Emotion related structures in large image databases. In: Proceedings of the ACM International Conference on Image and Video Retrieval, pp. 398–405. ACM, New York (2010)
17. Wang, W., He, Q.: A survey on emotional semantic image retrieval. In: International Conference on Image Processing (ICIP), pp. 117–120. IEEE, Los Alamitos (2008)
18. Wang, W., Yu, Y., Jiang, S.: Image retrieval by emotional semantics: A study of emotional space and feature extraction. In: International Conference on Systems, Man and Cybernetics (SMC 2006), vol. 4, pp. 3534–3539. IEEE, Los Alamitos (2006)
19. Wu, Q., Zhou, C., Wang, C.: Content-based affective image classification and retrieval using support vector machines. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 239–247. Springer, Heidelberg (2005)
20. Zhang, H., Koskela, M., Laaksonen, J.: Report on forms of enriched relevance feedback. Technical Report TKK-ICS-R10, Helsinki University of Technology, Department of Information and Computer Science (November 2008)
21. Zhang, H., Ruokolainen, T., Laaksonen, J., Hochleitner, C., Traunmüller, R.: Gaze- and speech-enhanced content-based image retrieval in image tagging. In: International Conference on Artificial Neural Networks–ICANN 2011, pp. 373–380 (2011)