Using Image Features and Eye Tracking Device to Predict Human Emotions Towards Abstract Images

Kitsuchart Pasupa[1(✉)], Panawee Chatkamjuncharoen[1], Chotiros Wuttilertdeshar[1], and Masanori Sugimoto[2]

[1] Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
kitsuchart@it.kmitl.ac.th, {panawee.c,chotiroswuttilertdeshar}@gmail.com
[2] Department of Computer Science, Hokkaido University, Sapporo 060-814, Japan
sugi@ist.hokudai.ac.jp

Abstract. Nowadays, emotional semantic image retrieval system enables users to access images that they want in a database according to emotional concept. This leads to affective image classification task which recently attracts researchers' attention. However, different users may experience different emotions depending on where, in the image, they are gazing on. This paper presents an improved prediction method by taking into account the users eye movement as implicit feedback while they are looking at the image. Our experimental results show that using both eye movement information and image feature together to determine users emotion gave more accurate predictions than using image feature alone.

Keywords: Eye movements · Implicit feedback · Emotion · Image retrieval

1 Introduction

Image retrieval is a method for users to search for their desired images in a database. In the past, each image in the database was manually tagged with keywords, then a system retrieved images by keyword matching. However, tagging keywords to a large number of images consumes a lot of resources. Therefore, less-resource-consuming retrieval methods were proposed i.e. automatic tag Image [1] and content-based image retrieval (CBIR). Images are described by lower-level features i.e. texture, color, and shape. CBIR compares features of a given image to every image in a database and then retrieves the images that have similar features to the given image. Recently, CBIR are becoming popular [2]. Examples of CBIR are PicSOM [3] and Google Image Search [4]. In 2006,

T. Bräunl et al. (Eds.): PSIVT 2015, LNCS 9431, pp. 419–430, 2016.
DOI: 10.1007/978-3-319-29451-3_34

Emotion Semantic Image Retrieval (ESIR) [5] was introduced - image search system with emotional concepts. From then on, the number of research studies on ESIR is steadily increasing. Examples of these studies are the following: a study of features of images that classify human emotions using theories of art and psychology [6]; a study of human emotions toward images of abstract art with low-level features [7]; and a study of emotion classification by using Multiple Kernel Learning (MKL) with basic image features – color, shape and texture [8].

In order to enhance the performance of CBIR, a relevance feedback from the user is used to refine the search results such as a mouse click. Clicking is an explicit feedback from a user actualizing his or her intention. A feedback can also be implicit. In 2009, Pasupa and his colleague developed a system that rank images by tracking and using the users unintentional eye movement to identify the ranking on images judged by users; specifically, the system outputs a ranking list of images that are likely to be the desired image [9]. However, the proposed approach is unrealistic and unable to apply to CBIR system because there is no eye movement presented a-priori on all images in database. Hence, Tensor Ranking Support Vector Machine is proposed to solve this problem [10]. In 2010, a CBIR system the so-called "PinView" was successfully developed to use both explicit and implicit feedback [11].

Emotional awareness depends on several factor such as gender, ethnicity, age and educational level. However, in this research, we were interested in using the close connection between human emotion and eye movement to classify the users emotion toward an image. The phrase "Beauty is in the eye of the beholder" leads to the assumption of this study that when several people look at a particular image, their emotions stimulated by the image may be different depending on which area of the image they are gazing on. This paper presents an approach to classify users emotions toward images by using both image feature and eye movement on the assumption that a better classification from using both together can achieve a higher prediction efficiency than using image feature alone.

The paper is organized as follows: Sect. 2 briefly explains theory used in this work. Data collection and feature extraction of this work is explained in Sects. 3 and 4, respectively. Section 5 shows experiment setup followed by its results and discussions in Sect. 6.

2 Theories

2.1 Emotion

Emotion is a feeling affected by stimulation. It can be classified by a dimension approach or a discrete approach.

The dimension approach considers each emotion as a combination of basic emotions that have their own intensities. Following this approach, Plutchik and Kellerman (1986) proposed eight basic emotions, which are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation [12]. These basic emotions can also be combined to create a derived emotion such as love, trust, etc.

On the other hand, the discrete approach considers that each emotion can have different intensities, as mentioned in [13,14]. In this work, we follow Ekman (1972) approach [15]. Six basic emotions were proposed which are anger, disgust, fear, happiness, sadness, and surprise. These six basic emotions are then grouped, in a study by Jack *et al.* (2014), by facial expressions into four categories which are happy, sad, angry, and fear [16]. Jack *et al.* (2014) also reported that fear and surprise have similar facial expressions as disgust and angry.

2.2 Abstract Art

Abstract art is the type of images that convey an artist's emotion or feeling as abstract forms. Wassily Kandinsky, the inventor of abstract art, had mentioned that the main elements of abstract art are colors and forms that are used to express meaning. Abstract art does not directly convey its meaning, thus individual viewers feeling and imagination are needed to interpret a piece of abstract art [17]. Emotions stimulated by a piece of abstract art are caused by each viewers interpretation of it which depends on where in the piece the viewer is gazing on. Therefore, even for the same image, each people can have a different emotion towards it. In this study, an eye tracker was used to record where in pieces of abstract art that viewers were gazing on. These records were then used to find the image features that can be used to classify the viewers emotion towards those pieces of abstract art.

2.3 Relevance Feedback

Relevance feedback is the information that a computer system gets from the user to determine whether the retrieved item is related to the users query or not. This piece of information is then used to optimize the response to the next query. There are two categories of relevance feedback: first, explicit – it is intentional feedback from users that takes time and mental effort such as mouse click, intentional eye movement, voice, and gesture; and second, implicit feedback – it is unintentional feedback that users do automatically. This type of feedback needs to be processed before use, such as heart rate, blood pressure, body temperature, and unintentional eye movement while reading a book or searching for information on a web browser.

2.4 Support Vector Machine

Support Vector Machine (SVM) is a well-known machine learning algorithm that is used to perform classification task. The are evidences that emotions towards images can also be classified by SVM [6–8,18]. SVM classifies data points by using linear hyperplane. Assuming that $\boldsymbol{x}_i$ is a sample vector, $\boldsymbol{x}_i$ pairs with y_i where $y \in \{-1, +1\}$, $\boldsymbol{w}$ is the weight vector, ξ is a non-negative slack variable, and C is a penalty constant. A hyperplane is generated such that the distance

between two groups of data is the farthest and the classification error is the lowest, as shown in (1).

$$\begin{aligned} &\min \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{s.t. } y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i - b) \geq 1 - \xi_i, \\ &\quad \xi \geq 0, i = 1\ldots m, \end{aligned} \tag{1}$$

where $C > 0$. The decision function is as follow:

$$f(\boldsymbol{x}) = \text{sgn}(\boldsymbol{w}\cdot\boldsymbol{x}_i - b) \tag{2}$$

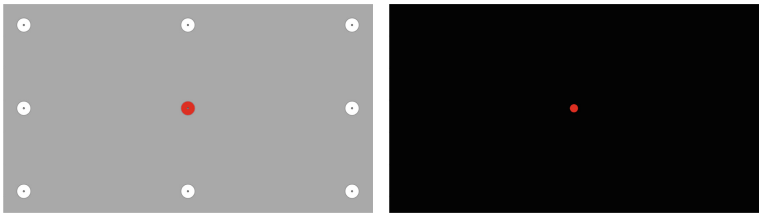
In real world application, a relationship between two groups of data is usually not linear in most cases, so a kernel trick is used. Examples of kernel functions are Polynomial and Radial Basis Function.

3 Data Collection

In this experiment, 100 sets of data of emotions towards images together with data of eye movements were collected from 20 undergraduate students aged 18–22 (10 males and 10 females) who had regular eyesight and were not wearing glasses. During data collection, our developed software displayed images from a database and collected subjects eye movement data with an eye tracker device at the same time. First, the subject was asked to sit in front of a 13-inch laptop (with a resolution of 1366×768 pixel) and then explained the procedure of the experiment. A subject had to specify his or her emotion towards a displayed image and an experiment operator recorded this information with a keyboard, as shown in Fig. 1. Before the experiment, the subjects were asked to try not to move their head, and the eye tracker was calibrated by having the subjects look at 9 points on the screen (1 point at the center and 8 points at the edges), as shown in Fig. 2a. Also, before skipping to the next image, the subjects were asked to look at an image that consists of a reference point at the center for 3 seconds, as shown in Fig. 2b, in order to make sure that his or her eyes were at the same position for every image.

The Eye Tribe Tracker could connect to a computer or a tablet and tracked eyes with data transfer rate of 30 Hz and $0.5° - 1°$ accuracy [19]. Abstract art images included in this study were obtained from Machajdik and Hanbury (2010) [6]. These included 228 images classified into 8 emotions – amusement, excitement, contentment, sad, anger, awe, fear, and disgust. However, we did not classify the images into 8 emotions as Wang *et al.* (2006) did, in order to reduce complexity for users to make judgements. Instead, we classified them according to the method of Shaver *et al.* (2001) [20]. Emotions are categorized to be 3 levels: primary, secondary, and tertiary emotion. These 8 emotions were categorized according to these levels of emotion into 4 primary emotions which are happy (amusement, excitement, contentment), sad, anger, and fear (fear, disgust), as in [16]. Images that produce awe emotion were not mentioned in [20], thus they were not considered in this study.

Fig. 1. The setup of eye tracking experiment.



(a) (b)

Fig. 2. (a) The nine-point calibration, and (b) the reference point screen.

Then, 100 images of 4 emotions (25 images for each emotion) that received the highest users votes – as in Machajdik and Hanbury (2010) [6] – were selected; some are shown in Fig. 3. Experimental results showed that when different people looked at the same image (Fig. 4a), their emotions were different. Users defined their emotion towards the image as fear, anger, and sad, as shown in Fig. 4b, c, and d, respectively.

4 Feature Extraction

For feature extraction in this study, we essentially extracted the color, shape, and texture features from images by using histograms. RGB was used to represent the distribution of color in each pixel of an image. The value of RGB is between 0–255. The shape feature used was as described by Sobel. It shows the magnitude of gradient in an image. The texture feature was extracted with a Gabor function with five scales and eight orientations. In this study, we used only one scale, eight orientations, and a high-pass filter. In order to increase the efficiency of emotional prediction, these features were combined by concatenating the vector of each feature together. These features were extracted from the original image, the original image with eye movement information, and the original image with eye movement information and Gaussian blur as shown in Fig. 5.

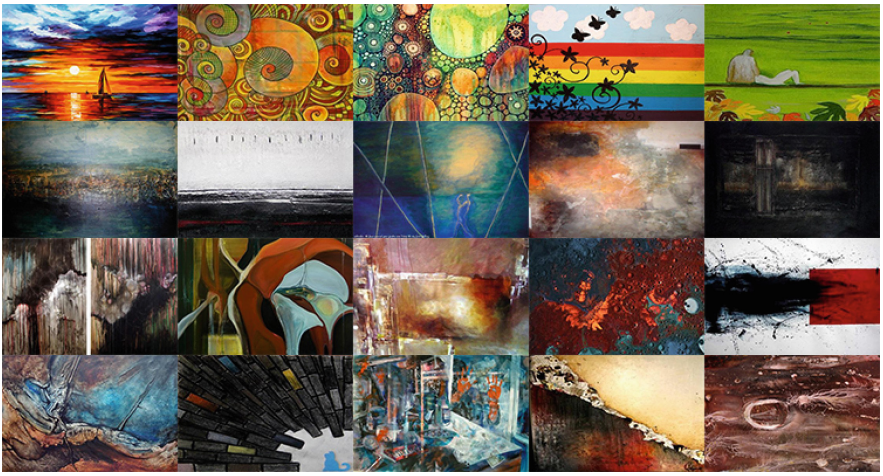
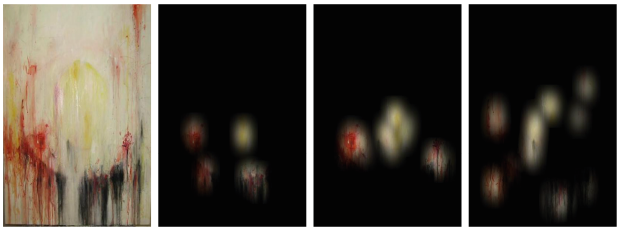
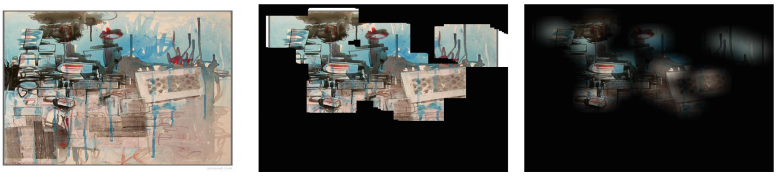


Fig. 3. Examples of images used in the study. Happy images are shown in 1st row followed by sad images in 2nd row. Anger and fear images are shown in 3rd, and 4th row, respectively.



(a) Original (b) Fear (c) Anger (d) Sad

Fig. 4. Different users with different emotions at the same image



(a) (b) (c)

Fig. 5. (a) Original Image, (b) Original image with eye movement information, and (c) Original image with eye movement information and Gaussian blur.

4.1 Original Image

Each original image were represented by the histograms of the distribution of RGB, Sobel and Gabor features at 8, 16, 32, and 64 bin sizes (n_{bin}). The size of color feature vector (n_{feature}) is n_{bin}^3 and the size of shape and texture feature vectors is n_{bin}.

4.2 Original Image Processed with Eye Movement Data

Original images were processed with eye movement data using a Gaussian function and the main cause for the change from each original image was which area in the image the subject was gazing on.

Mathematically, the processing can be described as follows. Let $I(x, y)$ be the image where (x, y) is a specified location and $E(x, y)$ is the eye movement information (fixation location). The Gaussian function that processes the image is expressed as

$$g(x_s) = e^{\frac{-x_s^2}{2\sigma}} \tag{3}$$

where x_s is a size of Gaussian. In this work, x_s is set to 100, 125, and 150 which gives Gaussian at 100×100, 125×125, and 150×150 sizes, respectively. σ is the standard deviation of the Gaussian function, here, we used $\sigma = \frac{x_s}{6}$. We used Gaussian function because it is well-suited for the foveation in human vision [21]. The Gaussian function transforms $E(x, y)$ into heatmap $G(x, y)$. Heatmap quantifies the degree of importance of part of image. Furthermore, $G(x, y)$ is be used to create a binary filter $H(x, y)$,

$$H(x, y) = \begin{cases} 1 \text{ if } & G(x, y) > 0 \\ 0 \text{ if } & G(x, y) = 0 \end{cases}. \tag{4}$$

Hence, the processed image is expressed as

$$I_E(x, y) = H(x, y) \times I(x, y) \tag{5}$$

where $\times$ denote element-wise multiplication.

4.3 Original Image Processed with Eye Movement Data and Gaussian Blur

Instead of using the filter $H(x, y)$ of Sect. 4.2, $G(x, y)$ is applied as the filter to process the original image into this type of processed image,

$$I_G(x, y) = G(x, y) \times I(x, y) \tag{6}$$

The values of $G(x, y)$ are between 0 and 1. If the value is close to 0, that position of the image is dark, whereas if the value is close to 1, that position is bright.

5 Experiment Setup

In this study, we performed two experiments, one that used each users model and the other one that used a global model, to compare the efficiencies of the 3 features – histograms of RGB colors, Sobel, and Gabor – in the three types of images – original image (I), original image processed with eye movement

data (I_E), and original image processed with eye movement data and Gaussian blur (I_G).

In the first experiment, three acquired features were compared by having them predicting each users emotion from his or her user model (20 users). SVM with linear kernel was used. The task is to classify 4 classes of data – happy, sad, anger, and fear, therefore, one-vs-all multi-class classification was considered. We used leave-one-out cross-validation (LOO-CV) technique to obtain the optimal model for each user based on misclassification rate. In this experiment, leave-one-image-out cross validation was applied to find the optimal n_{bin}, size of Gaussian filter, and SVM's C parameter. The adjustable C parameter is between 10^{-6} and 10^4. It should be noted that all data were normalized to zero mean and unit standard deviation.

In the second experiment, a global model constructed from all users data was used with similar setup as in the first experiment to predict the emotions of 10 users but with a leave-one-user-out cross-validation technique instead.

6 Experimental Results and Discussion

6.1 User Model

The effectiveness of emotion classification is reported as percentages of average accuracy of 20 users. The classification results of the three features and a set of four combined features on all three sets of data are compared to a baseline (BL) – that is, when SVM is able to predict only one emotion which is the majority class.

From Table 1, we can see that the average accuracies found from I, I_E, and I_G sets of data are better than that found from BL. Also, I_E and I_G are generally more effective than I but the texture feature of I is more effective than the two. Moreover, using combined features together is more effective than using a single feature alone. In combining them, the size of each feature is taken to be equal, a value in the range of {8, 16, 32, 64}-bin. We can see that the emotion prediction accuracy obtained from using all three features combined is the best at 48.05 %, while it is 47.60 % for combined color and texture features and 47.20 % for combined color and shape feature. However, using combined shape and texture features gives a worse prediction accuracy than using color feature alone. Therefore, we surmise that including color feature and eye movement in a set of combined features should increase prediction accuracy.

Figure 6 shows a graph of the averages of true positive rates from using every type of features, the most accurate prediction is from using the leftmost feature on the x-axis and the less accurate ones are from left to right (in that strict order for the I_E data set). It can be observed that the trend that a particular feature is more likely to be more or less accurate than the others is the same for the results reported in Table 1 and those reported in Fig. 6.

6.2 Global Model - New User

According to the first experiment, we found that I_E is more accurate than I_G. This might be because some parts of I_G is blurred and leads to information loss problem. Using the best combination of 3 features is more accurate than using any single feature or other combinations of features. Therefore, we chose to use I_E and the best combination of 3 features for predicting the emotions of any new users based on the leave-one-user-out cross-validated data on 10 users which were randomly sampled user.

Table 1. The accuracy of emotion predictions of 20 users (%) - user model.

Feature	BL	I	I_E	I_G
Color	36.35	43.35	**45.85**	45.70
Shape		41.95	**45.80**	44.80
Texture		**43.15**	42.20	42.15
Color+Shape		46.60	**47.20**	46.95
Color+Texture		45.65	**47.60**	45.65
Shape+Texture		41.40	**44.60**	44.45
Color+Shape+Texture		47.10	**48.05**	47.20
Average	36.35	44.17	**45.90**	45.26

From Tables 1 and 2, it can be observed that the average accuracies of the global model are higher than those of user model because the amount of the global models training data (900 images) is more than that of the user models training data (99 images). The results of the global model in Table 3 show that both I and I_E are more accurate than BL and that I_E is generally more accurate than I (55.70 % on average) except for the cases of the 5, 6, 7 subjects, that might be due to the differences in their eye movement behavior and their emotion towards an image to the majority of the subjects. Hence, predictions from the

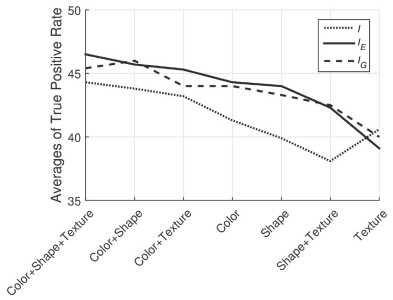


Fig. 6. Average of true positive predictions from using each feature.

Table 2. Average accuracies of the three contenders: I, I_E and BL on global model of 10 users.

User	BL	I	I_E
1	46.00	48.00	**50.00**
2	28.00	51.00	**55.00**
3	35.00	59.00	**65.00**
4	49.00	47.00	**57.00**
5	33.00	**49.00**	48.00
6	46.00	**62.00**	56.00
7	36.00	**55.00**	**55.00**
8	38.00	52.00	**53.00**
9	34.00	55.00	**61.00**
10	52.00	56.00	**57.00**
Average	39.70	53.40	**55.70**

Table 3. Average accuracies and their standard deviations of emotion prediction - global model.

n_{bin}	BL	I	I_E
8	39.70±7.96	51.90±4.23	**52.00±6.13**
16		52.50±5.46	**52.60±5.97**
32		51.30±5.14	**55.00±5.35**
64		50.70±5.14	**53.80±4.76**

Table 4. p-values of paired t-test of the accuracies of emotion prediction of the global model in Table 3.

n_{bin}	I-vs-BL	I_E-vs-BL	I_E-vs-I
8	0.001	0.001	0.955
16	0.002	0.001	0.949
32	0.005	0.001	0.034
64	0.006	0.001	0.001

global model with I_E for these exceptional subjects are less accurate than or equal to I. Moreover, comparing the accuracy of prediction of emotion in global model on the same bin size, I_E yields the best accuracy in all cases followed by I and BL as shown in Table 3.

Table 4 shows p-values of paired t-test of the prediction accuracies of emotion of the global model. It demonstrates that the results in Table 3 that I and I_E are significantly more accurate than BL for every bin size at $p < 0.01$ and that I_E are significantly more accurate than I only for 32, 64 bin size at $p < 0.05$.

Fig. 7. The average of true positive result for prediction of each emotion in global model.

Figure 7 shows the average of true positive results for each emotion and the average of true positive results for all emotions. It can be seen that I_E is more accurate than I in predicting sad, anger, and fear. Averaged over all emotions, I_E proves to be the best.

7 Conclusion

This paper presents an approach to predict users emotions towards abstract images (happy, sad, anger, and fear) by using (a) eye movement information and (b) low-level image features such as color, shape, and texture. It is found that using both the eye movement information and a set of image features gives more accurate predictions than using image features alone. The best combination of 3 features gives more accurate predictions than any other single feature or combinations. The color feature, either alone or in a combination, has the most positive influence on prediction results. Moreover, increasing the amount of training user data can also improve the prediction accuracy as shown in global model. However, eye movements can degrade the predictions if the users' judgements and eye movement behaviours are different from other users in the training model. Hence, user adaptation should be considered in the future.

Acknowledgments. The research leading to these results has received funding from the ASEAN University Network/Southeast Asia Engineering Education Development Network (AUN/SEED-Net) under the Short-term Research Program in Japan (SRJP 2014). This publication only reflects the authors views.

References

1. Ye, L., Ogunbona, P., Wang, J.: Image content annotation based on visual features. In: 8th IEEE International Symposium on Multimedia, pp. 62–69. IEEE Press, New York (2006)
2. Datta, R., Li, J., Wang, J.Z.: Content-based image retrieval: approaches and trends of the new age. In: 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 253–262. ACM (2005)

3. Laaksonen, J., Koskela, M., Oja, E.: PicSOM - self-organizing image retrieval with MPEG-7 content descriptors. IEEE Trans. Neural Netw. **13**, 841–853 (2002)
4. Google Image Search. https://images.google.com/
5. Wang, W.-N., Yu, Y.-L., Jiang, S.-M.: Image retrieval by emotional semantics: a study of emotional space and feature extraction. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 3534–3539. IEEE Press, New York (2006)
6. 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 (2010)
7. Zhang, H., Augilius, E., Honkela, T., Laaksonen, J., Gamper, H., Alene, H.: Analyzing emotional semantics of abstract art using low-level image features. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS, vol. 7014, pp. 413–423. Springer, Heidelberg (2011)
8. Zhang, H., Gönen, M., Yang, Z., Oja, E.: Predicting emotional states of images using Bayesian multiple kernel learning. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP 2013, Part III. LNCS, vol. 8228, pp. 274–282. Springer, Heidelberg (2013)
9. Pasupa, K., Saunders, C., Szedmak, S., Klami, A., Kaski, S., Gunn, S.: Learning to rank images from eye movements. In: 12th IEEE International Conference on Computer Vision Workshops, pp. 2009–2016. IEEE Press, New York (2009)
10. Hardoon, D.R., Pasupa, K.: Image ranking with implicit feedback from eye movements. In: 2010 Symposium on Eye-Tracking Research & Applications, pp. 291–298. ACM (2010)
11. Auer, P., Hussain, Z., Kaski, S., Klami, A., Kujala, J., Laaksonen, J., Leung, A.P., Pasupa, K., Shawe-Taylor, J.: PinView: implicit feedback in content-based image retrieval. In: Diethe, T., Cristianini, N., Shawe-Taylor, J. (eds.) Proceedings of the Workshop on Applications of Pattern Analysis 2010, vol. 11, pp. 51–57 (2010). Journal of Machine Learning Research
12. Plutchik, R., Kellerman, H.: Emotion: Theory, Research and Experience, vol. 3. Academic Press, New York (1986)
13. Izard, C.E.: Basic emotions, relations among emotions, and emotion-cognition relations. Psychol. Rev. **99**, 561–565 (1992)
14. Vytal, K., Hamann, S.: Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. J. Cogn. Neurosci. **22**, 2864–2885 (2010)
15. Ekman, P.: Universal and cultural differences in facial expression of emotion. In: Nebraska Symposium on Motivation, vol. 19, pp. 207–284. University of Nebraska Press, Lincoln (1972)
16. Jack, R., Garrod, O., Schyns, P.: Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Curr. Biol. **24**, 187–192 (2014)
17. Galenson, D.W.: Two paths to abstract art: Kandinsky and Malevich. Technical report 12403, National Bureau of Economic Research (2006)
18. Wu, Y., Bauckhage, C., Thurau, C.: The good, the bad, and the ugly: predicting aesthetic image labels. In: 20th International Conference on Pattern Recognition, pp. 1586–1589 (2010)
19. The Eye Tribe. https://theeyetribe.com
20. Shaver, P., Schwartz, J., Kirson, D., O'Connor, C.: Emotional knowledge: further exploration of a prototype approach. In: Parrott, G. (ed.) Emotions in Social Psychology: Essential Readings, pp. 26–56. Psychology Press, Philadelphia (2001)
21. Chang, E.C., Mallat, S., Yap, C.: Wavelet foveation. Appl. Comput. Harmon. Anal. **9**, 312–335 (2000)