Affective Abstract Image Classification and Retrieval Using Multiple Kernel Learning

He Zhang, Zhirong Yang, Mehmet Gönen, Markus Koskela, Jorma Laaksonen, Timo Honkela, and Erkki Oja

Department of Information and Computer Science
Aalto University School of Science, Espoo, Finland

{he.zhang,zhirong.yang,mehmet.gonen,markus.koskela,
jorma.laaksonen,timo.honkela,erkki.oja}@aalto.fi

Abstract. Emotional semantic image retrieval systems aim at incorporating the user's affective states for responding adequately to the user's interests. One challenge is to select features specific to image affect detection. Another challenge is to build effective learning models or classifiers to bridge the so-called "affective gap". In this work, we study the affective classification and retrieval of abstract images by applying multiple kernel learning framework. An image can be represented by different feature spaces and multiple kernel learning can utilize all these feature representations simultaneously (i.e., multiview learning), such that it jointly learns the feature representation weights and corresponding classifier in an intelligent manner. Our experimental results on two abstract image datasets demonstrate the advantage of the multiple kernel learning framework for image affect detection in terms of feature selection, classification performance, and interpretation.

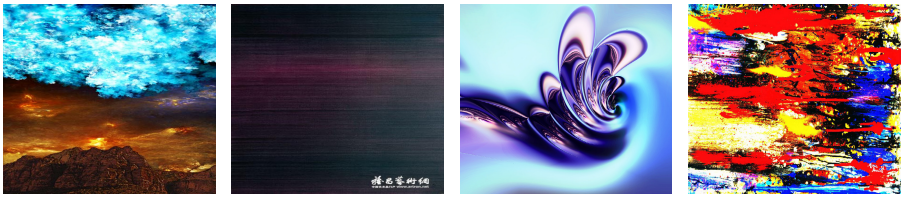
Keywords: Image affect, multiple kernel learning, group lasso, low-level image features, image classification and retrieval.

1 Introduction

Multimedia contents such as audio, image, and video contain information that can trigger people's affective feelings or emotions. Such information can be used by search engines for better modeling the user's preferences. Affective image classification and retrieval has attracted increasing research attention in recent years, due to the rapid expansion of the digital visual libraries on the Web. While most of the current content-based image retrieval (CBIR) systems [6] are designed for recognizing objects and scenes such as plants, animals, outdoor places etc., an emotional semantic image retrieval (ESIR) system [17] aims at incorporating the user's affective states to enable queries like "beautiful flowers", "cute dogs", "exciting games", etc.

Though emotions are highly subjective human factors, still they have certain stability and generality across different people and cultures [12]. As an example, Figure 1 shows four pictures taken from an abstract art image collection [19]. The

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 166–175, 2013.



(a) Exciting (b) Boring (c) Relaxing (d) Irritating

Fig. 1. Example images from the abstract art image data set [19] with the ground truth labels of Exciting, Boring, Relaxing, and Irritating

ground truth labels are determined by the emotion that has received the most votes from people. Intuitively, the "Exciting" and "Relaxing" pictures usually make people feel pleasant or evoke a positive feeling, whereas the "Boring" and "Irritating" pictures may evoke a negative feeling to the viewer.

In analogy to the concept of "semantic gap" that implies the limitations of image content description, the "affective 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" [8]. Among the challenges from image affect detection, one is to select suitable image features to reflect people's affective states, and another one is to build effective learning models or classifiers to bridge the "affective gap".

Many works (e.g., [5,11]) have focused on designing features specific to image affect detection, while others (e.g., [14,19]) simply utilized the traditional low-level color, shape, and texture features. Concerning the classifiers, support vector machines (SVM) [4] have been adopted in most of the works. However, one usually has to spend much time and effort in picking up the most suitable feature representation that can best reflect the viewer's emotions. For example, the authors in [14,19] utilized Fisher score to first rank and then select the most descriptive features, without considering the classifier at all. The authors in [5,11] picked each feature one by one with respectively an SVM and a naive Bayes classifier as the base learner to boost the performance, which requires explicit cross-validation steps for selecting features while optimizing the classifier parameters, and thus suffers from heavy computational complexities.

An image can be represented by different feature spaces. Multiple kernel learning (MKL) [2] can utilize all these feature representations simultaneously, such that it jointly learns the feature representation weights and the corresponding classifier for selecting automatically the most suitable feature representation or a combination of them. This can improve the classification performance and makes the interpretation of the results straightforward. MKL has earlier been applied for object detection in [16], and we are the first to introduce it into image affect detection. Our experimental results demonstrate the advantages of the MKL framework in affective classification and retrieval of abstract images.

Section 2 introduces the image features used in this paper. Section 3 introduces the MKL framework and an efficient algorithm that implements MKL. In Section 4, the experimental results on affective abstract image classification and retrieval are reported. Finally, the conclusions and future work are presented in Section 5.

2 Image Features

We have utilized a set of ten generic low-level color, shape, and texture features to represent each image. Table 1 gives a summary of these features. The features are extracted both globally and locally. For local features, a five-zone tiling mask is employed, where the image area is divided into four tiles by the two diagonals of the image, on top of which a circular center tile is overlaid [15]. All the features are extracted using the PicSOM system [10].

Table 1. The set of low-level image features used

Index	Feature	Type	Zoning	Dims.
F1	Scalable Color	Color	Global	256
F2	Dominant Color	Color	Global	6
F3	Color Layout	Color	8×8	12
F4	5Zone-Color	Color	5	15
F5	5Zone-Colm	Color	5	45
F6	Edge Histogram	Shape	4×4	80
F7	Edge Fourier	Shape	Global	128
F8	5Zone-Edgehist	Shape	5	20
F9	5Zone-Edgecoocc	Shape	5	80
F10	5Zone-Texture	Texture	5	40

Four of the features are standard MPEG-7 descriptors: Scalable Color, Dominant Color, Color Layout, and Edge Histogram. 5Zone-Color is defined as the average RGB values of all the pixels within the zone. 5Zone-Colm denotes the three central moments of HSV color distribution. Edge Fourier is calculated as the magnitude of the 16×16 FFT of Sobel edge image. 5Zone-Edgehist is the histogram of four Sobel edge directions. 5Zone-Edgecoocc is the co-occurrence matrix of four Sobel edge directions. Finally, 5Zone-Texture is defined as the histogram of the relative brightness of the neighboring pixels. More information about the features can be found in [15].

3 Multiple Kernel Learning

We can represent an image with different feature representations or views. However, the most suitable representation for a given task is generally not known a

priori. Instead of using a single representation (i.e., single-view learning), we can also make use of different representations simultaneously (i.e., multiview learning). Multiview learning with kernel-based methods is known as multiple kernel learning, which is a principled way of combining kernels calculated on different views to obtain a better prediction performance than single-view learning methods (see [7] for an extensive survey). In addition, MKL can learn the feature representation weights by itself according to the data and task at hand during the training stage without an explicit feature selection step, which makes the interpretation easy and straightforward.

Among the MKL algorithms, we use the group Lasso MKL [1] as our learning framework in that it is simple and efficient [18]. Both studies [1,18] have formulated an alternating optimization method that solves an SVM at each iteration and updates the kernel or feature representation weights η_m as follows:

$$\eta_m = \frac{\|w_m\|_2^{\frac{2}{p+1}}}{\left(\sum_{h=1}^{P} \|w_h\|_2^{\frac{2p}{p+1}}\right)^{\frac{1}{p}}} \tag{1}$$

where $\|w_m\|_2^2 = \eta_m^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k_m(x_i^m, x_j^m)$ is from the duality conditions. $k_m(-,-)$ denotes the kernel function calculated on the mth feature representation. P is the number of kernels or feature representations ($P = 10$ in our case), and p is chosen to be 1 so that $\sum_{m=1}^{P} \eta_m = 1$.

After updating the kernel weights in equation (1), the algorithm then solves a classical SVM problem by maximizing SVM dual formulation with the combined kernel $k = \sum_{m=1}^{P} \eta_m k_m$ as follows:

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{2}$$

subject to the constraints: $0 \leq \alpha_i \leq C$ for all $i = 1, ..., N$, and $\sum_{i=1}^{N} \alpha_i y_i = 0$, where C is the regularization parameter and y_i is the label (± 1) of training sample x_i. The two steps alternate until convergence.

4 Experiments

In this section, we present the experimental results using the MKL framework in the classification and retrieval of abstract images. We implemented the group Lasso MKL in MATLAB and took 20 alternating iterations for inference. We chose the LIBSVM [3] package for solving the classical SVM problem. For the group Lasso MKL, We set $C = 1$ and calculated the standard Gaussian kernel on each feature representation separately with the kernel width $s = 2\sqrt{D_m}$, where D_m is the dimensionality of corresponding feature representation. Therefore, no cross-validation steps are needed for learning the feature representation weights or the parameters of SVM classifier in group Lasso MKL.

4.1 Datasets

We have chosen abstract art images as our learning target instead of the photographic images since the latter contain contextual information that may affect the viewer's emotional assessment, which in turn would bias the learning results. Two abstract image datasets have been used in the experiments, `Abstract100`[1] [19] and `Abstract280`[2] [11].

The `Abstract100` dataset contains 100 images of abstract art paintings with different sizes and qualities through Google image search. These paintings were originally created by artists with various origins and periods. Each image has been evaluated by 20 college students (10 females and 10 males) including Asians and Europeans for two descriptive/adjective pairs "Exciting vs. Boring" and "Relaxing vs. Irritating" from the ratings of $\{-2, -1, 0, 1, 2\}$.

The `Abstract280` dataset contains 280 abstract art images that were peer-rated from a Web survey. Each image was labeled as the single emotion that had received the most votes from the eight affective categories: Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, and Sad(ness). The 280 images were rated by nearly 230 people, where each image was rated about 14 times.

4.2 Affective Abstract Image Classification

Experimental Setup. We use only the `Abstract100` dataset in this task, as SVM is optimized for binary classification problems. To obtain the ground truth labels for the classifier, we adopt a heuristic thresholding strategy: the image samples with ratings ≥ 0 in each descriptive pair are treated as the positive class, whereas those with ratings < 0 are treated as the negative class. For example, if an image receives an average rating of $(0.2, 1.5)$, then it is thresholded as $(+1, +1)$, which can be interpreted as both "Exciting" and "Relaxing". This results in roughly equal numbers of positive and negative samples. For training and testing, we use 5-fold cross-validation and calculate the average classification accuracy for each adjective pair. For comparisons, SVM_all uses the concatenation of all the 10 feature representations of an image as a single input, while SVM_best uses each of the 10 feature representations individually (as in [5,11]) and reports the one that has obtained the highest accuracy. Note that methods in both papers [5,11] require explicit cross-validation steps to select features and to optimize parameters (C and s), whereas no cross-validation procedures are involved in learning the adopted group Lasso MKL. The baseline result is calculated as the proportion of the majority class in each case.

Results. Figure 2 shows the average feature representation weights (i.e., kernel weights) in the range $[0, 1]$ based on 5-fold cross-validation using group Lasso MKL algorithm. We clearly see that, for the "Exciting vs. Boring" pair, Scalable Color (F1) ranks first, followed by Zone5-Color (F4), Edge-Histogram (F6), and

[1] An updated version: `http://research.ics.aalto.fi/cbir/abstract100`
[2] `http://www.imageemotion.org`

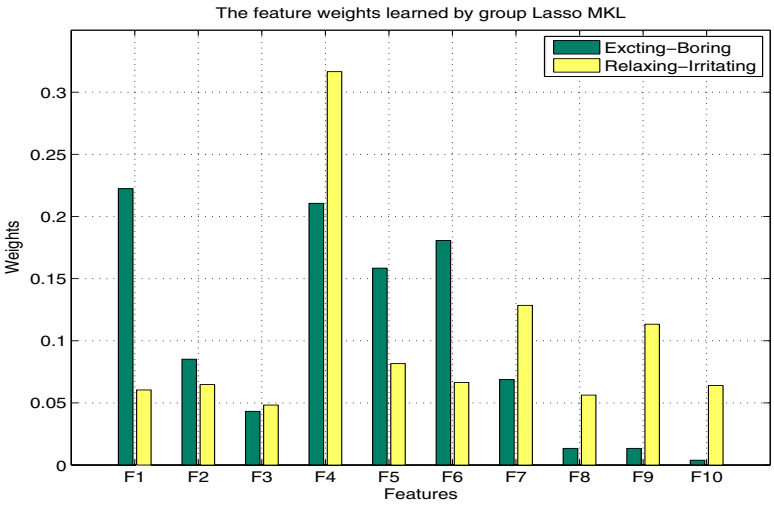


Fig. 2. The average feature representation weights over 5-fold cross-validation by using group Lasso MKL for two adjective pairs: Exciting-Boring and Relaxing-Irritating

Zone5-Colm (F5) etc. For the "Relaxing vs. Irritating" pair, Zone5-Color (F4) ranks first, followed by Edge Fourier (F7), Zone5-Edgecoocc (F9), and Zone5-Colm (F5) etc. This also confirms most of the studies (e.g., [11]) that colors and edges of an image are the most informative features for affect detection. Thus, multiple kernel learning serves as a natural testbed to identify the relative importance of feature representations automatically. Table 2 shows classification results on `Abstract100` dataset. It is clear that the group Lasso MKL

Table 2. The classification performances on `Abstract100` dataset. For SVM_all/SVM_best, we conducted grid search to choose the best (C, s) pair, with $C \in (0.5,\ 1,\ 2,\ 4,\ 8)$ and $s \in (0.0078,\ 0.0156,\ 0.0312,\ 0.0625,\ 0.1250,\ 0.25,\ 0.5,\ 1,\ 2)$.

Cases/Adjective Pair	Baseline	SVM_all	SVM_best	group Lasso MKL
Exciting-Boring	0.55	0.62	0.61	**0.67**
Relaxing-Irritating	0.55	0.55	0.72	**0.73**

Table 3. The computation time (s) of the comparison methods. All the methods were implemented in MATLAB on a Macintosh computer with an Intel Core i5 processor.

Cases/Adjective Pair	Baseline	SVM (all)	SVM (best)	group Lasso MKL
Exciting-Boring	–	6.10	9.70	**0.20**
Relaxing-Irritating	–	6.01	9.70	**0.20**

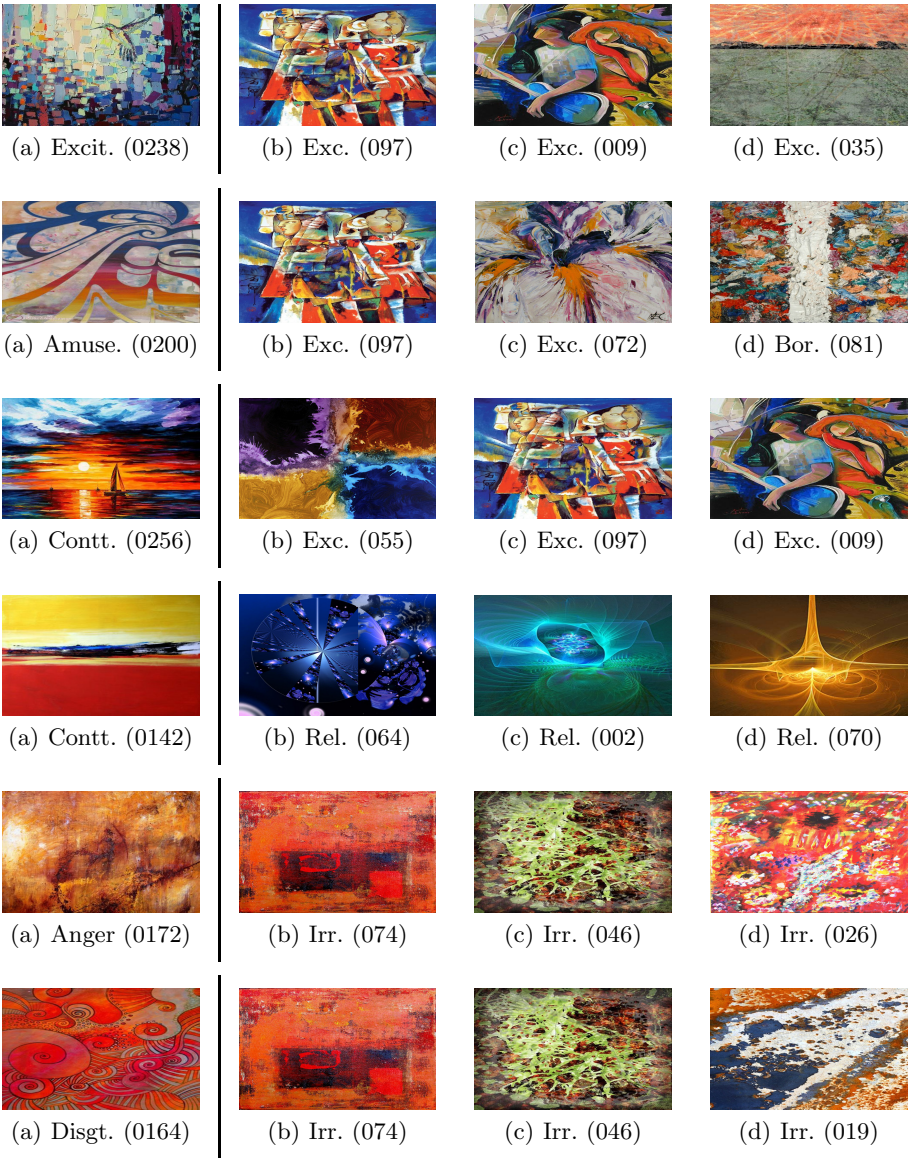


Fig. 3. The image retrieval results (displayed in "Groundtruth (index)" form) using the `Abstract280` images as queries shown in the first column, whereas the last three columns correspond to the top three retrieved images from the `Abstract100` dataset ranked by distance. The first three rows correspond to the query-retrieval results with kernel weights learned from Exciting-Boring adjective pair, whereas the last three rows correspond to the kernel weights of Relaxing-Irritating pair. Excit. = Excitement, Amuse. = Amusement, Contt. = Contentment, Disgt. = Disgust; Exc. = Exciting, Bor. = Boring, Rel. = Relaxing, Irr. = Irritating.

algorithm has achieved better classification performances than the other comparison methods in both cases. Table 3 gives the computation time of the compared methods. In either of the two cases, the computation time of group Lasso MKL is only about 1/30 of the SVM (all) and around 1/50 of the SVM (best).

4.3 Affective Abstract Image Retrieval

Experimental Setup. Both `Abstract100` and `Abstract280` datasets are used in this task. Firstly, we define the dissimilarity measure (the Euclidean distance in the implicit feature space) between a query image ($\boldsymbol{q}$) and a retrieved image ($\boldsymbol{r}$) as:

$$d_e(\boldsymbol{q},\boldsymbol{r}) = \sqrt{k_e(\boldsymbol{q},\boldsymbol{q}) + k_e(\boldsymbol{r},\boldsymbol{r}) - 2k_e(\boldsymbol{q},\boldsymbol{r})}$$

$$k_e(\boldsymbol{q},\boldsymbol{q}) = \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{q},\boldsymbol{q})$$

$$k_e(\boldsymbol{r},\boldsymbol{r}) = \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{r},\boldsymbol{r})$$

$$k_e(\boldsymbol{q},\boldsymbol{r}) = \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{q},\boldsymbol{r})$$

where $k_m(\cdot,\cdot)$ denotes the kernel function calculated on the mth feature representation and η_m is the weight for the corresponding kernel learned by the group Lasso MKL method. Therefore, given a query image $\boldsymbol{q}$, our aim is to find those images with the smallest $d_e(\boldsymbol{q},\boldsymbol{r})$ values. In essence, the smaller $d_e(\boldsymbol{q},\boldsymbol{r})$ is, the more probable that the retrieved image $\boldsymbol{r}$ evokes similar affective feelings in people. We use the `Abstract280` images as query images and let the MKL algorithm find the most relevant images from the `Abstract100` dataset. The kernel weights are selected on the complete set of `Abstract100` images (without splitting), either based on the "Exciting vs. Boring" or the "Relaxing vs. Irritating" adjective pair.

Results. Figure 3 shows the image retrieval results of certain query images for both cases. For the first case "Exciting vs. Boring", the "Excitement" image (0238) from `Abstract280` dataset successfully finds the other three "Exciting" images from `Abstract100` dataset as the first three returns. Similar results (except the "Boring" image (081)) can be observed for the "Amusement" image (0200) and the "Contentment" image (0256), due to the fact that the three emotional categories conceptually correlate with each other in the affective space [9]. For the second case "Relaxing vs. Irritating", the "Contentment" image (0142) also finds the other three "Relaxing" images as its top matches, which shows that an "Exciting" image often makes people feel "Relaxing" as well and vice versa. Both the "Anger" image (0172) and the "Disgust" image (0164) have retrieved "Irritating" images as their most relevant candidates. According to the Oxford Dictionary, the adjective word "Irritating" is defined as causing (someone) annoyance, impatience, anger, or irritation to a body part.

5 Conclusions and Future Work

In this paper, we have applied multiple kernel learning framework for affective classification and retrieval of abstract art images. MKL can make use of different feature representations or views of an image simultaneouly such that it jointly learns the feature representation weights and the corresponding classifier, which seeks for maximizing the classification performance without explicit feature selection steps. The group Lasso MKL algorithm has been adopted in the framework in that it is simple and efficient. The experimental results on two abstract image datasets have demonstrated the advantages of the group Lasso MKL in terms of feature selection, classification performance, and interpretation, for the affective abstract image classification and retrieval task.

It is worth emphasizing that MKL framework is not confined to detecting affect on abstract art images, but can be easily extended to other artistic (photographic) images and other affective stimuli such as audio and video data, given that the features and labels are available. Due to the varying subjectivity in humans and the limit of the available affective databases, it is of course not guaranteed that the MKL algorithm can make a perfect classification or retrieval for every single image. Methods such as zero-shot learning [13] may help to relieve the subjectivity and annotation issues. Eventually, the development in this interdisciplinary area relies on the joint efforts from, for instance, artificial intelligence, computer vision, cognitive science, psychology, and art theory.

Acknowledgements. This work has received funding from the Academy of Finland in the project Finnish Center of Excellence in Computational Inference Research (COIN). We gratefully acknowledge Ms Na Li for collecting the original `Abstract100` dataset.

References

1. Bach, F.: Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research 9, 1179–1225 (2008)
2. Bach, F., Lanckriet, G., Jordan, M.: Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning (ICML). ACM (2004)
3. Chang, C., Lin, C.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3) (2011)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
5. 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. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
6. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40(2), 5 (2008)
7. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of Machine Learning Research 12, 2211–2268 (2011)

8. Hanjalic, A.: Extracting moods from pictures and sounds: Towards truly personalized TV. IEEE Signal Processing Magazine 23(2), 90–100 (2006)
9. 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)
10. Laaksonen, J., Koskela, M., Oja, E.: PicSOM-self-organizing image retrieval with MPEG-7 content descriptors. IEEE Transactions on Neural Networks 13(4), 841–853 (2002)
11. 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)
12. 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)
13. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Advances in Neural Information Processing Systems (NIPS), pp. 1410–1418 (2009)
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(2) (2010)
15. Sjöberg, M., Muurinen, H., Laaksonen, J., Koskela, M.: PicSOM experiments in TRECVID 2006. In: Proceedings of the TRECVID 2006 Workshop (2006)
16. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proceedings of the 12th International Conference on Computer Vision (ICCV), pp. 606–613. IEEE (2009)
17. Wang, W., He, Q.: A survey on emotional semantic image retrieval. In: Proceedings of 15th IEEE International Conference on Image Processing, pp. 117–120 (2008)
18. Xu, Z., Jin, R., Yang, H., King, I., Lyu, M.: Simple and efficient multiple kernel learning by group lasso. In: Proceedings of the 27th International Conference on Machine Learning (ICML), pp. 1175–1182 (2010)
19. 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)