Content Based Detection of Popular Images in Large Image Databases

Martin Solli and Reiner Lenz

Media and Information Technology (MIT),
Department of Science and Technology (ITN), Linköping University,
SE-601 74 Norrköping, Sweden
{martin.solli,reiner.lenz}@liu.se
http://www.itn.liu.se/mit, http://diameter.itn.liu.se/

Abstract. We investigate the use of standard image descriptors and a supervised learning algorithm for estimating the popularity of images. The intended application is in large scale image search engines, where the proposed approach can enhance the user experience by improving the sorting of images in a retrieval result. Classification methods are trained and evaluated on real-world user statistics recorded by a major image search engine. The conclusion is that for many image categories, the combination of supervised learning algorithms and standard image descriptors results in useful popularity predictions.

1 Introduction

The motivation for this research is the basic need of every image search engine to show popular images in the search result, especially within the first images shown. In a typical real-life image retrieval task, the user queries a search interface with a keyword, for instance the name of an object. To satisfy the user the list of images that is returned should contain images of the desired object or scene. But what else makes an image popular? Here we investigate if ordinary image descriptors, together with supervised learning algorithms, can be used for estimating the popularity of images. The intended application is in large scale image search engines, where the proposed approach can improve the sorting of images in a retrieval result, and thereby enhancing the user experience. Either we can boost popular images, or do the opposite with non-popular images. We emphasize that in the current study we want to explore how far we can reach by using statistical measurements of image content only. The use of other relevance feedback tools, such as image click statistics, is not included in this paper. In a real-life application, however, the method can function as a complement to already implemented feedback methods. In the proposed approach we don't need to consider *why* an image is popular. Knowing that the image *is* popular is sufficient. We will train our system using two sub-sets of images, the most popular images, and remaining ones, in this paper referred to as non-popular images. Sub-sets are created based on recorded user behaviors in the Picsearch image search engine. Moreover, since the intended application is as a complement to

A. Heyden and F. Kahl (Eds.): SCIA 2011, LNCS 6688, pp. 218–227, 2011.

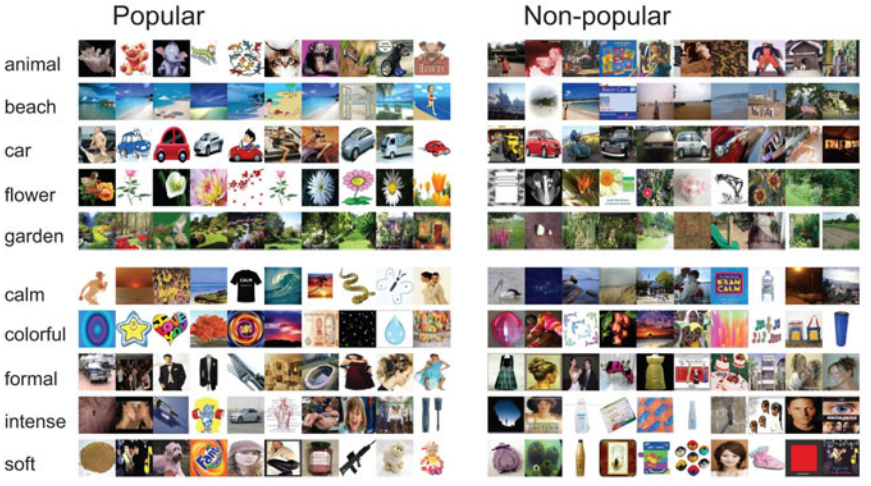


Fig. 1. Examples of popular and non-popular images from different keyword categories

other methods, we don't need to label every possible image. We will only label images that have a high probability, meaning that they are strong candidates for the popular or non-popular class.

Content based image retrieval has been an active research field for many years now. See for instance [12][3][9] for recent developments within image indexing and relevance feedback. The topic of estimating image popularity from statistical measurements of image content has not been addressed in the literature before. Instead we mention two papers by Datta et al. [1][2]. They use images from a photo sharing web page, peer-rated in two qualities, *aesthetics* and *originality*. Numerous visual or aesthetical image features, like Exposure, Depth-of-field, etc., are extracted. The relationship between features and observer ratings are explored through Support Vector Machines and classification trees, with the goal to build a model that can predict the quality of an image. A few other papers related to photo quality or aesthetics are Ke et al. [5] and Liu et al. [6][7]. We also mention Cohen-Or et al. [8] presenting a method that enhances the harmony among colors of a given image. There are however important differences between the references mentioned above, and the work presented here. First, since the standard approach for displaying image retrieval results is to display image thumbnails, we will only work with small images (maximum size 128 pixels). Earlier work typically use images of much larger size. Secondly, for many methods predicting photo quality, numerous specialized image descriptors are developed. Here we prefer to start our investigations using common image descriptors, with the advantage that they are already computed in many image retrieval systems.

2 The Image Database

Our database is collected from the image search engine[1] belonging to Picsearch AB (publ). The database contains thumbnail images, with a maximum size of 128 pixels (height or width), together with meta-data, such as keywords/labels and user statistics. Original images were crawled from public web pages using 20 different keywords, given in Table 1. 10 of them are related to ordinary objects, and 10 are based on emotional properties. Image thumbnails were shown to users visiting the Picsearch search engine, and statistics of how many times each image has been viewed and clicked were recorded. The ratio "number of clicks / number of views" is used as an estimate of popularity, but only for images that have been viewed at least 50 times. For each image category we start by splitting the images into two sub-groups, the 1000 most popular, and remaining ones. We sample 100 images from the remaining ones, and save them as non-popular images. As popular images we save the 100 most popular images. In other words, each category will be described by 200 images, 100 popular, and 100 non-popular. To illustrate the database, the 10 most popular, and 10 non-popular images, for examples of categories, are plotted in Fig. 1. Each image category in our database typically contains several thousand images, so it may sound strange that only 100+100 images are used from each category. The reason is that the popularity score declines quite rapidly for many image categories, making it risky to include more than 100 images in the popular class.

Table 1. The keywords used in the experiments. 1-10 are representing objects (or scenes), and 11-21 are related to emotions.

1:	animal	5:	garden	9:	food	13:	formal	17:	cold
2:	beach	6:	cat	10:	lion	14:	intense	18:	warm
3:	car	7:	dog	11:	calm	15:	soft	19:	pure
4:	flower	8:	doll	12:	colorful	16:	vivid	20:	quiet

3 Image Descriptors

There is a huge number of image descriptors that can be applied in the following experiments. However since a comprehensive comparison of image descriptors is beyond the scope of this study, we will limit ourselves to two local and two global image descriptors. As local descriptors we use bag-of-features (or bag-of-words) models, known as state-of-the-art solutions in object and scene classification. These are compared to global histogram descriptors. The descriptors are:

RGB-histogram: 512 ($8 \times 8 \times 8$) bins RGB-histogram, with equally sized bins.

[1] http://www.picsearch.com/

Bags-of-emotions: A color-based emotion-related image descriptor proposed by Solli and Lenz [15]. The descriptor is based on an emotion metric derived in psychophysical experiments, and the assumption that color emotions in images are mainly affected by homogenous regions, and transitions between regions. Emotion scores are derived for found regions, and transitions between regions, and values are saved in a *bag-of-emotions*, which is a 112 bins histogram. The result is a single histogram, and not a collection of histograms as in ordinary bag-of-features models.

SIFT: Scale Invariant Feature Transform, a standard tool in image processing and computer vision, proposed by Lowe [13]. We use a SIFT implementation by Andrea Vedaldi[2], both for interest point detection, and descriptor extraction. The result is a 128 bins histogram describing each interest point.

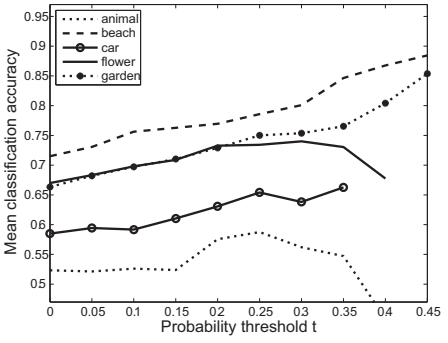
OpponentSIFT: The recommended descriptor in the evaluation of color descriptors carried out by van de Sande et al. [14]. In OpponentSIFT, all three channels in the opponent color space are described by the SIFT descriptor. One of the channels contains the intensity information, whereas the others contain color information invariant to changes in light intensity. The descriptor is included in a software package by van de Sande[3].

In the following we will refer to the above descriptors as "RGB", "ebags", "SIFT" and "OSIFT". The average number of found interest points per thumbnail (for SIFT/OSIFT) is 125, which is believed to be sufficient for the intended application. For SIFT and OSIFT, we adopt the common procedure for bag-of-features models and perform clustering in the descriptor space (also known as codebook generation), followed by vector quantization to obtain the distribution over cluster centers (the distribution over codewords). For clustering we use k-means, with 500 cluster centers, and 10 iterations, each with a new set of initial centroids. Then we search for the iteration that returns the minimum within-cluster sums of point-to-centroid distances. Clustering is carried out with 10 000 descriptors (1000 descriptors randomly selected from each of the keywords 1-5 and 11-15). State-of-the-art solutions in image classification are often using codebooks of even greater size. But since our experiments focus on thumbnail images, where the number of found interest points is relatively low, we find it appropriate to limit the size to 500 cluster centers. Preliminary experiments with an increased codebook size did not result in increased performance. Similar conclusions about the size of the codebook can for instance be found in van Gemert et al. [4]. The ebags histogram has 112 dimensions (bins), whereas SIFT and OSIFT have 500 (after vector quantization), and RGB has 512 bins. For an easier and fair comparison, we use Principal Component Analysis to reduce the number of dimensions of the RGB, SIFT and OSIFT histograms, leaving the 112 dimensions with the highest variance.

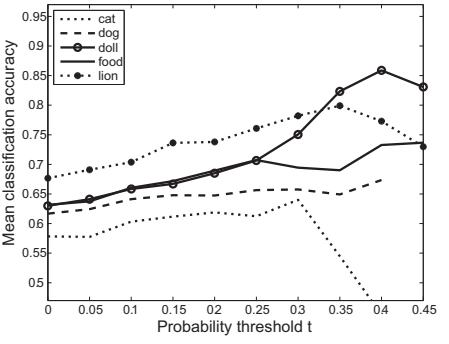
[2] http://www.vlfeat.org/∼vedaldi/
[3] http://www.colordescriptors.com/

Table 2. The overall classification accuracy for different descriptors, different sets of images (objects and/or emotions), and two different values on the threshold t

t	Images	RGB	ebags	SIFT	OSIFT	mean
0	all	0.57	0.58	0.55	0.56	0.57
0	objects	0.60	0.61	0.58	0.61	0.60
0	emotions	0.55	0.56	0.53	0.54	0.55
0.25	all	0.52	0.70	0.69	0.71	0.66
0.25	objects	0.68	0.77	0.72	0.74	0.73
0.25	emotions	0.43	0.58	0.50	0.65	0.54
mean		0.56	0.63	0.60	0.64	

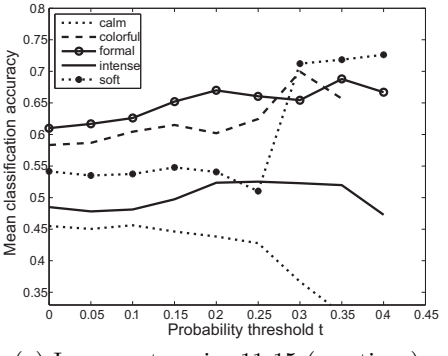


(a) Image categories 1-5 (objects)

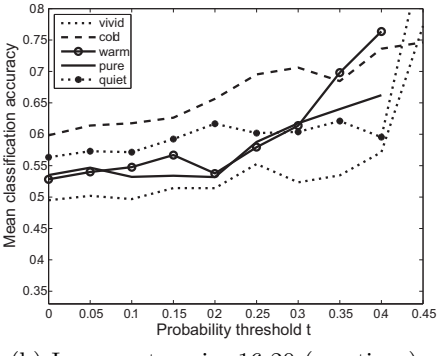


(b) Image categories 6-10 (objects)

Fig. 2. The mean classification accuracy over descriptors RGB, ebags, and OSIFT, for different object categories and varying values of t



(a) Image categories 11-15 (emotions)



(b) Image categories 16-20 (emotions)

Fig. 3. The mean classification accuracy over descriptors RGB, ebags, and OSIFT, for different emotion categories and varying values of t

4 Classification

With database images separated into popular and non-popular images, the goal is to be able to predict what class an unknown image belongs to. Our classification experiments are based on a stratified 10-fold cross-validation procedure. The original image set of 200 images belonging to each category is partitioned into 10 subsets, each containing the same number of popular and non-popular images (10 + 10). The cross-validation process is repeated K times, where $K-1$ subsets are used for training the classifier, and the remaining subset is used for validating the model. After K training runs, each image in the category has received a classification score, and obtained scores are used for deriving the overall classification accuracy. An advantage with this kind of cross-validation is that all images are used for both training and validation, which is useful when the number of images is limited. A disadvantage, however, is that we obtain 10 classification models for each category. Depending on the final application, we might need to combine the result from all models.

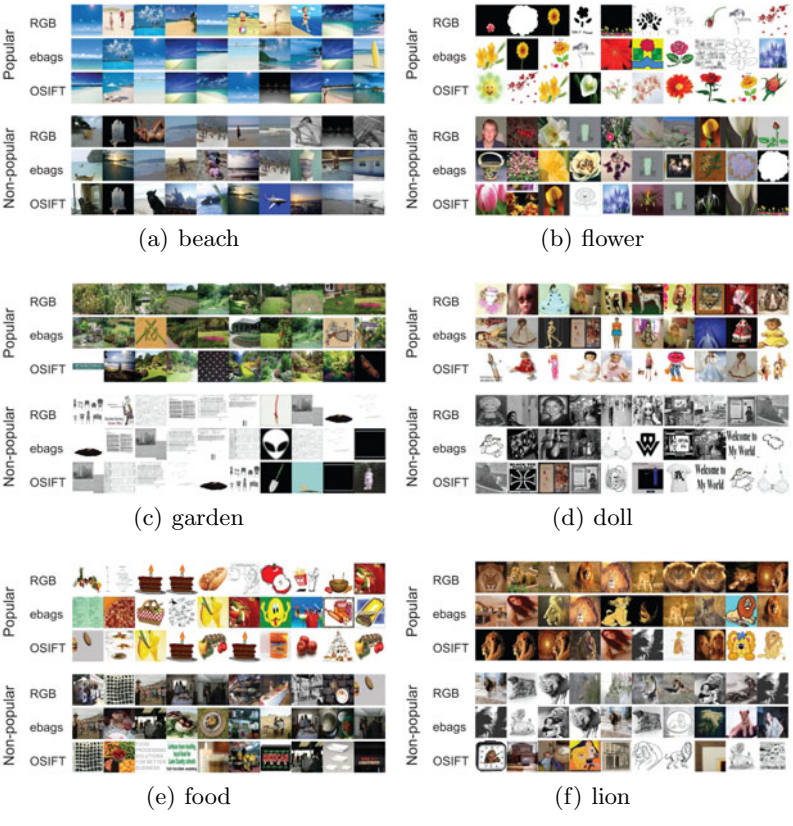
A common method for solving a two-class problem is to utilize a supervised learning algorithm, for instance a Support Vector Machine. Here we use *SVM-light* by Thorsten Joachims [10]. For simplicity, and to ensure reproducibility, all experiments are carried out with default settings. Obtained classification scores are translated to probabilities using the method proposed by Lin et al. [11]. For the intended application, it is not crucial that every single image is labeled with popular or non-popular. As an alternative, we only label images that have a probability estimate close to 1 or 0, meaning that they are strong candidates for the popular and non-popular class respectively. For image i, with probability estimate p_i, we define a probability threshold t. Image i will only be classified if p_i lies outside the interval $\{0.5 - t \leq p_i \leq 0.5 + t\}$.

5 Results

The classification accuracy for different descriptors, different sets of images, and two different thresholds t, can be seen in Table 2. Here the classification model was trained and tested on merged image sets, containing images from all emotion categories, or all object categories, or a large set containing both emotions and objects. The classification accuracy is given by the proportion of correctly labeled images (e.g. 0.8 means that 80% of the images were labeled correctly). Obtained scores indicate that it is harder to predict popularity in emotion categories than in object categories. Since the SIFT descriptor usually performs poorer than OSIFT, we will exclude the SIFT descriptor from the remaining experiments. The overall performance for all descriptors is rather poor (an accuracy close to 0.5 is equivalent to a random classification). When we, however apply the learning algorithm on individual categories, we notice large differences in accuracy between different categories. The result for different object categories can be seen in Fig. 2. The plot shows the relationship between the mean classification accuracy over descriptors RGB, ebags and OSIFT, and different

Table 3. The classification accuracy for the best performing image categories, and three different descriptors. (t=0.3)

Image category	RGB	ebags	OSIFT
beach	0.82	0.82	0.77
flower	0.77	0.76	0.69
garden	0.80	0.73	0.72
doll	0.78	0.70	0.77
food	0.70	0.72	0.67
lion	0.80	0.81	0.74
colorful	0.68	0.70	0.72
formal	0.76	0.57	0.63
soft	1.00	0.61	0.53
cold	0.72	0.71	0.69
warm	0.72	0.63	0.49
mean	0.78	0.70	0.67

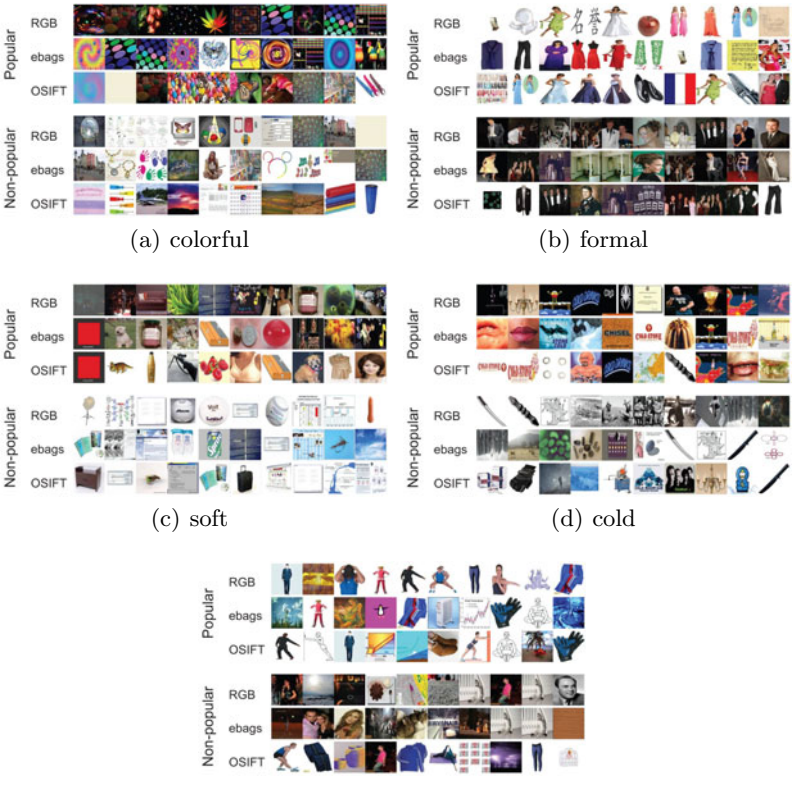


(a) beach (b) flower

(c) garden (d) doll

(e) food (f) lion

Fig. 4. Classification examples for the best performing object categories



(a) colorful (b) formal

(c) soft (d) cold

(e) warm

Fig. 5. Classification examples for the best performing emotion categories

values on the probability threshold t. Similar plots for the emotion categories can be seen in Fig. 3. For some categories, a raised probability threshold eventually results in an empty class, shown as a terminated curve in the figure. We see that for object categories *beach*, *flower*, *garden*, *doll*, *food* and *lion*, the accuracy is relatively high, and shows consistency for an increased t value. The same holds for the emotion categories *colorful*, *formal*, *soft*, *cold* and *warm*. Remaining categories, especially the remaining emotion categories, show poor performance. The classification accuracy for the best performing categories, for different image descriptors, can be seen in Table 3. We find a value of $t = 0.3$ appropriate. The RGB histogram performs best, followed by ebags and OSIFT.

We illustrate the classification result by plotting examples of classified images. For each image category, and the descriptors RGB, ebags and OSIFT, the 10 images that obtained the highest probability score (most popular) are plotted together with the 10 images that obtained the lowest score. Plots for the object categories *beach*, *flower*, *garden*, *doll*, *food* and *lion* are shown in Fig. 4, and plots for the emotion categories *colorful*, *formal*, *soft*, *cold* and *warm* are shown in Fig. 5. As we might expect, for some of the categories, especially *cold* and

Table 4. The mean classification accuracy (over RGB, ebags, OSIFT) for different categories and different image subsets: only popular, or only non-popular images. The table also shows the number of images belonging to each subset. (t=0.3)

	Accuracy (Nr of images)	
Image category	popular	non-popular
beach	0.79 (57)	0.82 (54)
flower	0.72 (41)	0.77 (39)
garden	0.63 (39)	0.84 (38)
doll	0.62 (33)	0.84 (41)
food	0.67 (35)	0.72 (34)
lion	0.82 (43)	0.74 (44)
colorful	0.70 (25)	0.72 (23)
formal	0.72 (37)	0.43 (22)
soft	0.69 (18)	0.39 (20)
cold	0.67 (33)	0.74 (34)
warm	0.60 (27)	0.64 (23)

warm, the popularity of the images seem to have very little in common with emotional color properties.

In our final experiments we derive the mean classification accuracy (over descriptors RGB, ebags and OSIFT) for subsets containing popular and non-popular images only. The result is shown in Table 4, including the average number of images that were classified to belong to each subset. Subsets are rather small due to the threshold t (here $t = 0.3$). However, since many users only look at the first few images in a search result, a popular subset of only 20-30 images is often sufficient. And depending on the application, we can of course use the probability estimate to rank all images in a category, not only the popular or non-popular ones. We notice that it is usually easier to classify non-popular images than popular ones (even if it is completely the opposite in a few categories), but there is no general relationship between the classification accuracy and the number of images in each subset.

6 Summary and Conclusions

We have investigated the use of standard image descriptors, both local and global, for estimating the popularity of thumbnail images. The intended application is in large scale image search engines, where the estimate of popularity can be used (in conjunction with other methods) to improve the ordering of images in a retrieval result, and thereby enhancing the user experience. The topic is crucial for any large scale image search engine. In the experiments, a Support Vector Machine was used in a 10-fold cross-validation procedure to distinguish between popular and non-popular images. To our surprise, the best performing descriptor is a global descriptor, the traditional RGB histogram, followed by Bags-of-emotions, and OpponentSIFT. The classification accuracy,

however, varies significantly between different image categories. In the current experiments, the popularity estimate was proven to be useful in 11 out of 20 image categories. By using earlier recorded user statistics for individual image categories, one can easily decide which categories the proposed approach can be applied to. The overall conclusion is that for many image categories, the combination of supervised learning algorithms and standard image descriptors results in useful popularity predictions. An advantage of using standard descriptors is that these are often already included in many image databases. The next step would be to explore how to combine the result with other types of features, for instance real-time user feedback based on image click statistics.

References

1. 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)
2. Datta, R., Li, J., Wang, J.: Learning the consensus on visual quality for next-generation image management. In: 15th ACM Int. Conf. on Multimedia, MM 2007, Augsburg, pp. 533–536 (2007)
3. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2) (2008)
4. van Gemert, J.C., Veenman, C.J., Smeulders, A.W., Geusebroek, J.-M.: Visual word ambiguity. IEEE TPAMI 32, 1271–1283 (2010)
5. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 419–426 (2006)
6. Liu, L., Chen, R., Wolf, L., Cohen-Or, D.: Optimizing photo composition. Computer Graphics Forum (Proceedings of Eurographics) 29, 469–478 (2010)
7. Liu, L., Jin, Y., Wu, Q.: Realtime aesthetic image retargeting. In: Proc. of Eurographics WS on Computational Aesthetic in Graphics, Visualization, and Imaging, pp. 1–8 (2010)
8. Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.-Q.: Color harmonization. In: ACM SIGGRAPH 2006, vol. 25, pp. 624–630 (2006)
9. Huang, T.S., Dagli, C.K., Rajaram, S., Chang, E.Y.: Active Learning for Interactive Multimedia Retrieval. Proceedings of the IEEE 96(4), 648–667 (2008)
10. Joachims, T.: Making large-scale support vector machine learning practical. In: Advances in Kernel Methods: Support Vector Learning, pp. 169–184 (1999)
11. Lin, H.T., Lin, C.J., Weng, R.C.: A note on platt's probabilistic outputs for support vector machines. Mach. Learn. 68(3), 267–276 (2007)
12. Liu, Y., Zhang, D., Lu, G., Ma, W.-Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recogn. 40(1), 262–282 (2007)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
14. van de Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. IEEE TPAMI 32, 1582–1596 (2010)
15. Solli, M., Lenz, R.: Color based bags-of-emotions. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 573–580. Springer, Heidelberg (2009)