A Selective Weighted Late Fusion for Visual Concept Recognition

Ningning Liu, Emmanuel Dellandrea, Chao Zhu,
Charles-Edmond Bichot, and Liming Chen

Université de Lyon, CNRS,
Ecole Centrale de Lyon, LIRIS, UMR5205, F-69622, France
{ningning.liu,emmanuel.dellandrea,chao.zhu,
charles-edmond.bichot,liming.chen}@ec-lyon.fr

Abstract. We propose in this paper a novel multimodal approach to automatically predict the visual concepts of images through an effective fusion of visual and textual features. It relies on a Selective Weighted Late Fusion (SWLF) scheme which, in optimizing an overall Mean interpolated Average Precision (MiAP), learns to automatically select and weight the best experts for each visual concept to be recognized. Experiments were conducted on the MIR Flickr image collection within the ImageCLEF 2011 Photo Annotation challenge. The results have brought to the fore the effectiveness of SWLF as it achieved a MiAP of 43.69 % for the detection of the 99 visual concepts which ranked 2^{nd} out of the 79 submitted runs, while our new variant of SWLF allows to reach a MiAP of 43.93 %.

Keywords: Visual concept recognition, multimodality, feature fusion.

1 Introduction

Machine-based recognition of visual concepts aims at automatically recognizing high-level visual semantic concepts (HLSC), including scenes (e.g., indoor, outdoor, landscape, *etc.*), objects (car, animal, person, *etc.*), events (travel, work, *etc.*), or even emotions (melancholic, happy, *etc.*). It proves to be extremely challenging because of large intra-class variations and inter-class similarities, clutter, occlusion and pose changes. The past decade has witnessed tremendous efforts from the research communities as testified the multiple challenges in the field, e.g., Pascal VOC [1], TRECVID [2] and ImageCLEF [3]. Most approaches to visual concept recognition (VCR) have so far focused on appropriate visual content description, and have featured a dominant bag-of-visual-words (BoVW) representation along with local SIFT descriptors. Meanwhile, increasing works in literature have discovered the wealth of semantic meanings conveyed by the abundant textual captions associated with images [4]. Therefore, multimodal approaches are proposed for VCR by making joint use of user textual tags and visual descriptions to bridge the gap between HLSC and low-level visual features. The work presented in this paper is in that line and targets an effective feature fusion scheme for VCR.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 426–435, 2012.

As far as multimodal approaches are concerned, it requires a fusion strategy to combine information from multiple sources, *e.g.*, visual stream and sound stream for video analysis [5], textual and visual content for multimedia information retrieval [6], *etc.* This fusion can be carried out at feature level [7], namely *early fusion*, or at score level [8], *i.e. late fusion*, or even at some intermediate level, *e.g.*, such as kernel level [9]. While early fusion is straightforward and simply consists of concatenating the features extracted from various information sources into a single representation, its disadvantage is also well known: the curse of dimensionality and the difficulty in combining features of different natures into a common homogeneous representation. As a result, late fusion strategies, which consist of integrating the scores as delivered by the classifiers on various features through a fixed combination rule, *e.g.*, sum, are the preferred fusion method in literature [10,11]. They not only provide a trade-off between preservation of information and computational efficiency but also consistently yield better performance as compared to early fusion methods [5]. Furthermore, a comprehensive comparative study of various combination rules, *e.g.*, sum, product, max, min, median, and majority voting, by Kittler *et al.* [12], suggests that the sum rule is much less sensitive to the error of individual classifiers when estimating posterior class probability.

The proposed fusion scheme, namely Selective Weighted Late Fusion (SWLF), falls into the category of late fusion strategies. Specifically, when different features, *e.g.*, visual ones and textual ones, can be used for VCR, SWLF learns to automatically select and weight the best experts to be fused for each visual concept to be recognized. The proposed SWLF builds on two simple insights. First, the score delivered by a feature, *i.e.* expert, should be weighted by its intrinsic quality for the classification problem at hand. Second, in a multi-label scenario where several visual concepts may be assigned to an image, different visual concepts may require different features which best recognize them. For instance, the "sky" concept may greatly require global color descriptors, while the best feature to recognize a concept like street could be a segment-based feature for capturing straight lines of buildings. Furthermore, we also propose three different variants of SWLF which are compared using data provided by the ImageCLEF 2011 photo annotation task. The experimental results demonstrate the effectiveness of the proposed approach.

The rest of this paper is organized as follows. The proposed fusion scheme, SWLF, is presented in section 2. The experiments we have conducted to evaluate SWLF are described in section 3. Finally, we conclude and give some insight on the future work in section 4.

2 Selective Weighted Late Fusion

The proposed SWLF scheme has a learning phase which requires a training dataset for the selection of the best experts and their corresponding weights for each visual concept. Specifically, given a training dataset, we divide it into two disjoint parts composed of a training set and a validation set. For each visual concept, a binary classifier (concept versus no concept) is trained, which is also

called *expert* in the subsequent, for each type of features using the data in the training set. Thus, for each concept, we generate as many experts as the number of different types of features. The quality of each expert can then be evaluated through a quality metric using the data in the validation set. In this work, the quality metric is chosen to be the interpolated Average Precision (iAP). The higher iAP is for a given expert, the more weight should be given to the score delivered by that expert for the late fusion. Concretely, given a visual concept k, the quality metrics, *i.e.* iAP, produced by all the experts are first normalized into w_k^i. To perform a late fusion of all these experts at score level, the *sum of weighted scores* is then computed as in (1):

$$score : z_k = \sum_{i=1}^{N} (w_k^i * y_k^i), \tag{1}$$

where y_k^i represents the score of the i^{th} expert for the concept k, and w_k^i stands for the normalized iAP performance of the feature f_i on the validation dataset. In the subsequent, late fusion through (1) is called *weighted score* rule.

For the purpose of comparison, we also consider three other score level fusion schemes, namely "min", "max" or "mean" respectively expressed as $min : z_k = min(y_k^1, y_k^2, ..., y_k^N)$, $max : z_k = max(y_k^1, y_k^2, ..., y_k^N)$, $mean : z_k = \frac{1}{N} \sum_{i=1}^{N} y_k^i$.

Actually, these three fusion rules can have very simple interpretation. The *min* fusion rule is the consensus voting. A visual concept is recognized only if all the experts recognize it. The *max* rule can be called alternative voting. A visual concept is recognized as long as one expert has recognized it. Finally, the *mean* rule can be assimilated as the majority voting where a concept is recognized if the majority of the experts recognize it.

In practice, one discovers that the late fusion of all the experts leads to a decrease in the global classification accuracy, *i.e.* the mean iAP over the whole set of visual concepts. The reason could be that some of features so far proposed can be noisy and irrelevant to a certain number of visual concepts, thus disturbing the learning process and lowering the generalization skill of the learnt expert on the unseen data. For this purpose, we further implement the SWLF scheme inspired by a wrapper feature selection method, namely the SFS method (Sequential Forward Selection) [13], which firstly initializes an empty set, and at each step the feature that gives the highest correct classification rate along with the features already included is added to the set of selected experts to be fused. More specifically, for each visual concept, all the experts are sorted in a decreasing order according to their iAP. At a given iteration N, the only first N experts are used for late fusion and their performance is evaluated over the data of the validation set. N is increased until the overall classification accuracy measured in terms of MiAP starts to decrease.

2.1 The Learning Algorithm of SWLF

The learning procedure of the SWLF algorithm can be defined as follows:

Selective Weighted Late Fusion (SWLF) algorithm for training

Input: Training dataset T (of size N_T) and validation dataset V (of size N_V).
Output: Set of N experts for the K concepts $\{C_k^n\}$ and the corresponding set of weights $\{\omega_k^n\}$ with $n \in [1, N]$ and $k \in [1, K]$.
Initialization: $N = 1$, $MiAP_{max} = 0$.

- Extract M types of features from T and V
- For each concept $k = 1$ to K
  - For each type of feature $i = 1$ to M
    1. Train the expert C_k^i using T
    2. Compute ω_k^i as the iAP of C_k^i using V
  - Sort the ω_k^i in descending order and denote the order as j^1, j^2, ..., j^M to form $W_k = \{\omega_k^{j^1}, \omega_k^{j^2}, ..., \omega_k^{j^M}\}$ and the corresponding set of experts $E_k = \{C_k^{j^1}, C_k^{j^2}, ..., C_k^{j^M}\}$
- For the number of experts $n = 2$ to M
  - For each concept $k = 1$ to K
    1. Select the first n experts from E_k : $E_k^n = \{C_k^1, C_k^2, ..., C_k^n\}$
    2. Select the first n weights from W_k : $W_k^n = \{\omega_k^1, \omega_k^2, ..., \omega_k^n\}$
    3. For $j = 1$ to n : Normalise $\omega_k^{j'} = \omega_k^j / \sum_{i=1}^n \omega_k^i$
    4. Combine the first n experts into a fused expert, using the *weighted score* rule through (1): $z_k = \sum_{j=1}^n \omega_k^{j'}.y_k^j$ where y_k^j is the output of C_k^j
    5. Compute $MiAP_k^n$ of the fused expert on the validation set V
  - Compute $MiAP = 1/K. \sum_{k=1}^K MiAP_k^n$
  - If $MiAP > MiAP_{max}$
    - Then $MiAP_{max} = MiAP$, $N = n$
    - Else break

2.2 The Variants of SWLF

As the number of experts N is the same for each concept in the above algorithm, this version of SWLF is called *SWLF_FN* (fixed N). However, several variants can be built upon SWLF. Indeed, instead of fixing the same number of experts N for all concepts, it is possible to select the number of experts on a per-concept basis. Thus the number of experts can be different for each concept. We have also implemented this variant denoted *SWLF_VN* (variable N) in the following. Another variant concerns the way the experts are selected at each iteration. Instead of adding the n^{th} best expert at iteration n to the set of previously selected $n-1$ experts, one can also select the expert which yields the best combination of n experts, in terms of $MiAP$, once added to the set of $n-1$ experts already selected at the previous iteration. This variant is denoted *SWLF_SFS* in the following as the selection scheme is inspired from the feature selection method "Sequential Feature Selection" [13] generally used for early fusion of features.

3 Experiments

In order to allow a comparison of our method with those among the most recent ones in the visual concept recognition domain, we carried out experiments on the MIR Flickr image collection that was used within the ImageCLEF 2011 Photo Annotation Challenge [3]. The goal of this challenge was to automatically annotate images according to 99 visual concepts. The database is a subset of MIRFLICKR-1M image collection from thousands of real world users under a creative common license. It is split into a training set of 8,000 images and a test set of 10,000 images. Each image is provided with a textual description (user tags).

The measure we have considered to evaluate the classification performance is the Mean interpolated Average Precision (MiAP) that is also used in ImageCLEF 2011 Photo Annotation Challenge.

3.1 The Features

More and more, images are provided with textual resources such as Exif data, legends, tags. These data can be easily obtained on the sharing websites such as Flickr[1], which is the data source of the MIRFLICKR-1M, and the textual descriptions are a rich source of semantic information that is interesting to consider for the purpose of image classification and retrieval.

Therefore, in order to describe images for further classification, we propose to use not only visual features extracted from the image, but also textual features extracted from the textual resources associated with images. These features are briefly presented in the next subsections.

Visual Features. As the concepts to be detected in images can be characterized by many different visual properties, we extract a rich set of features. Indeed, we consider low-level features based on color, texture, shape, being local or global, as well as mid-level features related to aesthetic and affective image properties. The color features are moments and histograms computed with several different color spaces such as RGB and HSV. The texture features are based on cooccurrences [14] and on different variants of Local Binary Patterns (LBP) using several scales and color spaces [15]. Shape feature are histograms of image line orientations extracted from Hough transform [16]. Several variants of SIFT features are also extracted using a dense grid and different color spaces [17]. Among the mid-level features, we extract aesthetic features proposed in [18] and [19] as well as affective features related to color harmony and dynamism [20].

In total, we extract 24 visual feature sets of various dimensions ranging from 1 for color harmony to 4000 for each of the SIFT variants (the size of the codebook).

Textual Features. The textual resources associated with images can take many forms. We consider in this paper that images are provided with a set of words

[1] `http://www.flickr.com/`

(or tags), as it is the case with the MIR Flickr image collection that we use in our experiments.

Our goal here is to extract a semantic information from this text. To do so, we use a feature that is defined as a histogram of textual concepts towards a vocabulary or dictionary where each bin of this histogram represents a concept of the dictionary, whereas its value is the accumulation of the contribution of each word within the text document toward the underlying concept according to a semantic similarity measure provided by Wordnet ontology [21]. For instance, the bin associated with the concept "rain" of the dictionary can be activated by tags such as "water", "liquid", "precipitation", "dripping liquid", "monsoon", *etc.*.

As several dictionaries and semantic similarity measures are conceivable, we extract 10 variants of this textual histogram feature leading to a total of 10 textual features.

3.2 Experimental Setup

The initial training dataset, provided by ImageClef 2011 for the photo annotation challenge, was first divided into a training set (50%, 4005 images) and a validation set (50%, 3995 images), and balanced the positive samples of most concepts as half for training and half for validation. The proposed features, both textual and visual, were then extracted from the training and validation sets. Support Vector Machines (SVM) [22] were chosen as classifiers (or experts) for their effectiveness both in terms of computation complexity and classification accuracy. A SVM expert was trained for each concept and each type of features, as described in section 2. Following J. Zhang *et al.* [23], we used χ^2 kernel for histogram-based features and RBF kernels for the other features. The RBF and χ^2 kernel functions are defined by: $K_{rbf}(F, F') = exp^{-\frac{1}{2\sigma^2}\|(F-F')\|^2}$ and $K_{\chi^2}(F, F') = exp^{\frac{1}{I}\sum_{1=1}^{n}\frac{(F_i-F_i')^2}{F_i+F_i'}}$ where F and F' are the feature vectors, n is their size, I is the parameter for normalizing the distances which was set at the average value of the training set, and σ was set at $\sqrt{n/2}$.

We made use of the LibSVM library [24] as SVM implementation (C-Support Vector Classification). The tuning of the different parameters for each SVM expert was performed empirically according to our experiments, in which the weight of negative class ("-w-1") was set at 1, and the weight for positive class ("-w1") was optimized on the validation set using a range of 1 through 30.

In the following, we give results of SWLF on the validation set which is a part of the training set provided for the ImageCLEF 2011 challenge, but also on the test set on which participants were evaluated as it has been released after the competition.

3.3 Experimental Results

The fusion schemes we propose in this paper have been used to combine the 24 visual and 10 textual features presented in section 3.1, and applied to MIR Flickr

image collection. Figure 1 presents the MiAP performance achieved by the basic SWLF scheme, SWLF_FN, using the "score" rule for combining experts which is compared with the standard fusion operators "min", "max" and "mean". These results are given on both the validation and test sets and show the evolution of the MiAP as N, the number of features to be fused, is increased from 1 to 34.

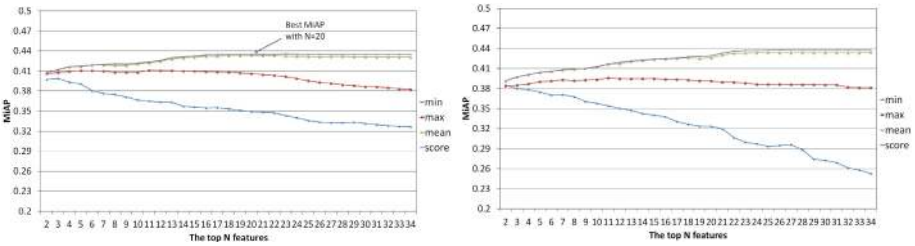

Fig. 1. The MiAP performance of SWLF_FN using different rules ("min", "max", "mean" and "score") for fusing visual and textual features using the validation set (a) and the test set (b).

As we can see from Figure 1 (a), the max and min-based SWLF_FN schemes tend to decrease the MiAP when the number of features to be fused, N, is successively increased from 1 to 34. On the contrary, the performance of weighted score and mean-based SWLF_FN schemes keep increasing until N reaches 20 and then stays stable. These results demonstrate that the weighted score and mean-based SWLF schemes perform consistently better than the max and min-based fusion rules. While close to each other, the weighted score-based SWLF_FN scheme performs slightly better than the mean-based SWLF_FN scheme. Figure 1 (b) presents the results obtained using the test set. We can observe that the results are very close to those obtained using the validation set, which proves the very good generalization skill of SWLF_FN, particularly when using "mean" and "score" fusion rules.

Table 1. The MiAP obtained by SWLF_FN, SWLF_VN and SWLF_SFS on the validation and test sets

Method	MiAP on the validation set	MiAP on the test set
SWLF_FN(N=20)	43.55 %	42.71 %
SWLF_FN(N=22)	43.53 %	43.69 %
SWLF_VN	44.51 %	38.61 %
SWLF_SFS	44.03 %	43.93 %

A comparison of the MiAP obtained by the three SWLF variants (SWLF_FN, SWLF_VN and SWLF_SFS) is provided in Table 1. It confirms the good generalization skill of SWLF since the MiAP obtained on the test set is very similar as

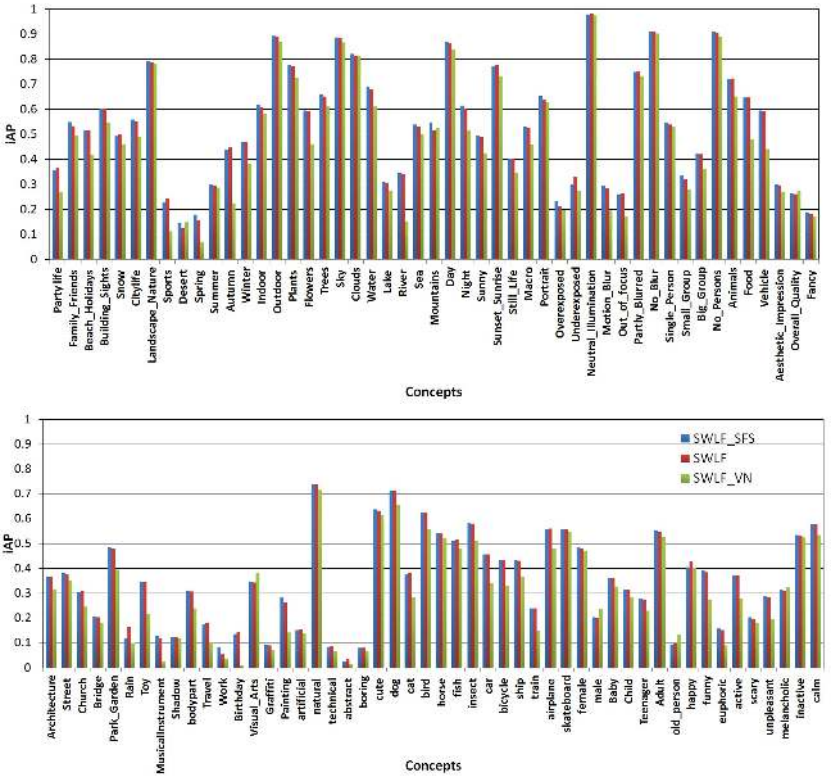

Fig. 2. The iAP obtained by SWLF_FN, SWLF_VN and SWLF_SFS for the 99 concepts of ImageCLEF 2011 Photo Annotation challenge

the one obtained on the validation set. The best result is obtained by SWLF_SFS with a MiAP of 43.93 % on the test set, closely followed by SWLF_FN (with $N = 22$) with a MiAP of 43.69 %. SWLF_VN is the least efficient among SWLF variants. Indeed, although it performs slightly better than SWLF_FN and SWLF_SFS on the validation set, its performance drops by more than 5 % on the test set. This tends to suggest that SWLF_VN, in optimizing the iAP on a per class-basis, is more prone to overfitting than SWLF_FN and SWLF_SFS, thus leading to a more severe performance drop on unseen data (test dataset).

Figure 2 presents the iAP obtained by SWLF_FN, SWLF_VN and SWLF_SFS for each of the 99 concepts that had to be detected within the ImageCLEF 2011 Photo Annotation challenge. One can notice that the slight superiority of SWLF_SFS over SWLF_FN based on the global MiAP is respected for most of the concepts, as well as the lower results obtained by SWLF_VN. This Figure also shows that some concepts are very well detected such as "Neutral_Illumination", "Outdoor", "Sky" with an iAP around 90 % whereas some are very difficult to detect such as "Abstract", "Boring", "Work" with an iAP lower than 10 %.

The results of the runs submitted by the different teams participating to the ImageCLEF 2011 Photo Annotation challenge are reported in [3]. The best results are obtained by teams using multimodal approaches (visual and textual features). The first rank was obtained by TUBFI with a MiAP of 44.3 % followed by our submission using SWLF_FN with a MiAP of 43.7 % (this is given in Table 2 of [3]). This result has been improved after our participation thanks to the proposition of SWLF_SFS which displays a MiAP of 43.9 %, proving its effectiveness for combining visual and textual features.

4 Conclusion

We have presented in this paper a novel Selective Weighted Late Fusion (SWLF) that iteratively selects the best features and weights the corresponding scores for each concept at hand to be classified. Three variants of SWLF, namely SWLF_FN, SWLF_VN and SWLF_SFS, have been proposed and compared.

Experiments were conducted on the image collection within the ImageCLEF 2011 Photo Annotation challenge. Our submission using SWLF_FN obtained a MiAP of 43.69 % for the detection of the 99 visual concepts which ranked 2^{nd} out of the 79 submitted runs. This results has even been improved by the variant SWLF_SFS developed after our participation, reaching a MiAP of 43.93%.

The experimental results have also shown that SWLF, in efficiently fusing visual and textual features, displays a very good generalization ability on unseen data for the image annotation task with a multi-label scenario.

Acknowledgments. This work is partly supported by the french ANR under the project VideoSense ANR-09-CORD-026.

References

1. Everingham, M., Van Gool, L.J., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision, 303–338 (2010)
2. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR 2006: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330 (2006)
3. Nowak, S., Nagel, K., Liebetrau, J.: The clef 2011 photo annotation and concept-based retrieval tasks. In: CLEF Workshop Notebook Paper (2011)
4. Guillaumin, M., Verbeek, J.J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: CVPR, pp. 902–909 (2010)
5. Snoek, C.G.M., Worring, M., Smeulders, A.W.M.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 399–402 (2005)
6. Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., Renders, J.M.: Crossing textual and visual content in different application scenarios. Multimedia Tools and Applications 42, 31–56 (2009)
7. Snoek, C.G.M., Worring, M., Geusebroek, J.M., Koelma, D.C., Seinstra, F.J.: The mediamill trecvid 2004 semantic video search engine. In: Proceedings of the TRECVID Workshop (2004)

8. Westerveld, T., Vries, A.P.D., van Ballegooij, A., de Jong, F., Hiemstra, D.: A probabilistic multimedia retrieval model and its evaluation. EURASIP Journal on Applied Signal Processing 2003, 186–198 (2003)
9. Binder, A., Samek, W., Kloft, M., Müller, C., Müller, K.R., Kawanabe, M.: The joint submission of the tu berlin and fraunhofer first (tubfi) to the imageclef2011 photo annotation task. In: CLEF Workshop Notebook Paper (2011)
10. Wu, Y., Chang, E.Y., Chang, K.C.C., Smith, J.R.: Optimal multimodal fusion for multimedia data analysis. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, pp. 572–579 (2004)
11. Znaidia, A., Borgne, H.L., Popescu, A.: Cea list's participation to visual concept detection task of imageclef 2011. In: CLEF Workshop Notebook Paper (2011)
12. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20, 226–239 (1998)
13. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recogn. Lett. 15, 1119–1125 (1994)
14. Haralick, R.M.: Statistical and structural approaches to texture. Proceedings of the IEEE 67, 786–804 (1979)
15. Zhu, C., Bichot, C.E., Chen, L.: Multi-scale color local binary patterns for visual object classes recognition. In: ICPR, pp. 3065–3068 (2010)
16. Pujol, A., Chen, L.: Line segment based edge feature using hough transform. In: The Seventh IASTED International Conference on Visualization, Imaging and Image Processing, VIIP 2007, pp. 201–206 (2007)
17. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1582–1596 (2010)
18. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: CVPR, vol. 1, pp. 419–426 (June 2006)
19. Datta, R., Li, J., Wang, J.Z.: Content-based image retrieval: approaches and trends of the new age. In: Multimedia Information Retrieval, pp. 253–262 (2005)
20. Dellandréa, E., Liu, N., Chen, L.: Classification of affective semantics in images based on discrete and dimensional models of emotions. In: International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 99–104 (June 2010)
21. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM 38, 39–41 (1995)
22. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer New York Inc., New York (1995)
23. Zhang, J., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vision 73, 213–238 (2007)
24. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 1–27 (2011)