Fine-Grained Image Classification Using Color Exemplar Classifiers

Chunjie Zhang[1], Wei Xiong[1], Jing Liu[2], Yifan Zhang[2], Chao Liang[3], and Qingming Huang[1,4]

[1] School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049, Beijing, China
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
[3] National Engineering Research Center for Multimedia Software, Wuhan University, 430072, Wuhan, China
[4] Key Lab of Intell. Info. Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
{cjzhang,wxiong,qmhuang}@jdl.ac.cn, {jliu,yfzhang}@nlpr.ia.ac.cn, cliang@whu.edu.cn

Abstract. The use of local features has demonstrated its effectiveness for many visual applications. However, local features are often extracted with gray images. This ignores the useful information within different color channels which eventually hinders the final performance, especially for fine-grained image classification. Besides, the semantic information of local features is too weak to be applied for high-level visual applications. To cope with these problems, in this paper, we propose a novel fine-grained image classification method by using color exemplar classifiers. For each image, we first decompose it into multiple color channels to take advantage of the color information. For each color channel, we represent each image with a response histogram which is generated by exemplar classifiers. Experiments on several public image datasets demonstrate the effectiveness of the proposed color exemplar classifier based image classification method.

Keywords: Color space, exemplar classifier, fine-grained image classification, structured regularization.

1 Introduction

Recently, fine-grained image classification becomes more and more popular, such as flower classification [1, 2] and bird classification [3]. The goal of fine-grained image classification is to separate images of a basic-level category. Although related with generic image classification, fine-grained classification pays more attention to separate highly similar images which are with subtle differences [4]. The traditional bag-of-visual-words (BoW) model often fails to solve this problem. This is because gray image is used for local feature extraction which ignores the discriminative information within different color spaces. For example, to classify flower images in

B. Huet et al. (Eds.): PCM 2013, LNCS 8294, pp. 327–336, 2013.

Fig. 1. The white daisy can be easily separated with yellow dandelion by color. However, it is relatively more difficult to separate them by using shape features such as SIFT.

Figure 1, the white daisy can be easily separated with yellow dandelion by color. However, it is very difficult to separate them by using local shape features such as SIFT. The same thing also happens when classifying different birds [3].

To combine the color information, researchers have proposed many color based features [5-8]. Weijer *et al.* [5] analyzed the error propagation to the hue transformation and weighted the hue histogram with its saturation. To combine the statistical information of different colors, Mindru *et al.* [6] defined the generalized color moments. Bosch *et al.* [7] used the HSV color model and extracted SIFT features over the three channels which results in a 3×128 dimensional vector. However, this vector has no invariance properties which limit its discriminative power. The OpponentSIFT is generated in a similar way by Sande *et al.* [8] by transforming images into the opponent color space. One problem with these color based descriptors is that they treated the color space jointly without considering the differences within different color spaces. We believe this information should be used to boost the classification performance.

Besides, local features carry little semantic information which is often not discriminative enough for visual classification. To generate more semantically meaningful representation, many works have been done [9-14]. Some researchers [9-12] tried to generate a semantic space with all the training images. Rasiwasia and Vasconcelos [9] proposed a holistic context models for visual recognition. By defining each class as the group of database images labeled with a common semantic label, Carneiro *et al.* [10] proposed a probabilistic formulation for image annotation and retrieval which is conceptually simple and do not require prior semantic segmentation of images. Yao *et al.* [11] proposed a discriminative feature mining method and trained random forest with discriminative decision trees to incorporate the region semantic information and achieved good results. Li *et al.* [12] proposed the Object Bank which represented an image as a response map of generic object detectors. Other researchers [13-15] made use of exemplar image separately. Sparsity constraint [16] is also used by researchers. Malisiewicz *et al.* [13] proposed to ensemble exemplar SVMs for object detection and achieved comparable results with the state-of-the-art methods. Zhang *et al.* [14] used supervised semantic representation for scene classification. Although proven effective, these methods also ignored the color information. If we can combine this semantic based image representation with color information, we will be able to representation images more discriminatively and eventually improve the image classification performance.

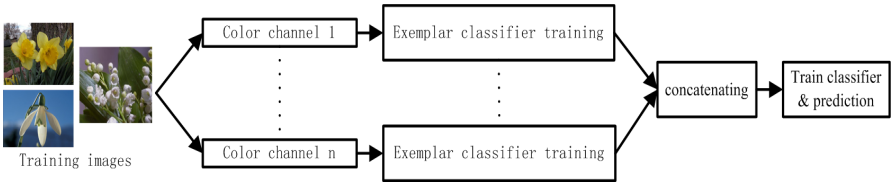

Fig. 2. Flowchart of the proposed method

In this paper, we propose a novel fine-grained image classification method using the color exemplar classifiers. For each channel, we first decompose each image into different color channels (*e.g.* RGB, HSV). Then we extract SIFT local features and use the sparse coding along with max pooling technique to represent images of this channel. Exemplar SVM classifiers are trained and we use the outputs of each exemplar classifier as the final image representation for this channel. Each image is represented by concatenating the final image representation of all color channels. Experimental results on several public datasets demonstrate the effectiveness of the proposed method. Figure 2 gives the flowchart of the proposed method.

The rest of this paper is organized as follows. Section 2 gives the details of the proposed color exemplar classifier based fine-grained image classification method. We give the experimental results and analysis in Section 3. Finally, we conclude in Section 4.

2 Color Exemplar Classifier Based Fine-Grained Image Classification

In this section, we give the details of the proposed color exemplar classifier based fine-grained image classification algorithm. We first decompose images into multiple color channels. For each channel, exemplar classifiers are trained and we use the output of these exemplar classifiers for image representation. We then concatenate the image representation for each channel as the final image representation. Finally, we learn one-vs-all SVM classifiers to predict the categories of images.

2.1 Color Channel Decomposition and Local Feature Extraction

The color information plays an important role for visual application. The traditional BoW model often extracts local features on gray images without considering the color information. Besides, the state-of-the-art color descriptors [5, 8] jointly consider the color spaces without making good use of the information within each color channel. Figure 3 shows an example of this problem. Different color channels review different aspects of the images and should be treated separately.

To make full use of the color information as well as take advantage of different color channels, we propose to first decompose images into different color channels, such as RGB, HSV. For each channel, we extracted SIFT features and use these features to represent images. This is different from traditional color descriptors [5, 8]

which concatenate the local features of all the color channels to form a long descriptor. We treat each channel separately to preserve the discriminative information of each color channel. Besides, to represent images for each channel, we use the popular sparse coding with max pooling strategy [17] as it has been proven very effective. In this way, we are able to combine the color information for better image representation which will eventually help to improve the final image classification performance.

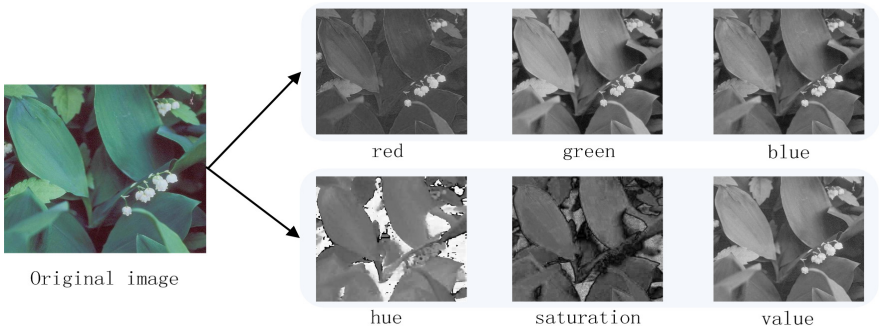

Fig. 3. The decomposition of one image into different color channels (RGB, HSV). It is best viewed in color.

2.2 Image Classification with Color Exemplar Classifier

After representing images for each channel, we are able to take advantage of this color representation. However, the local feature carries little semantic information. To get a more semantic image representation, we use the exemplar classifier based method. For each color channel, we train exemplar classifier for each image and use the output of these exemplar classifiers as the new image representation. We choose linear SVM as the exemplar classifier. This is because it is very effective and computational efficient.

Formally, let $p_{i,j}^{k} \in \mathbb{R}^{M\times 1}, i=1,2,...,N, j=1,2,..., M, k=1,2,...,K$ be the predicted values of the i-th image with the j-th class of the k-th color channel, where N is the number of training images and M is the number of training image categories, K is the number of color channels. For color channel k, the exemplar classifier based image representation is $h_i^k = [h_{i,1}^k;...;h_{i,j}^k;...;h_{i,M}^k]$. To combine the image representation of multiple color channels, we concatenate the h_i^k for each channel k, where $k=1,2,...,K$, as $h_i = [h_i^1, h_i^2, ..., h_i^K]$ for the i-th image.

To predict the categories of images, we use the one-vs-all strategy and train linear SVM classifiers. Let $y_i \in Y = \{1,2,...,M\}$ be the corresponding labels of image i, we try to learn M linear functions such that:

Fig. 4. Example images of the Oxford Flower 17 dataset

$$y = \max_{c \in Y} w_c^T h \tag{1}$$

This is achieved by solving the following optimization problem as:

$$\min_{w_c} \|w_c\|^2 + \lambda \sum_{i=1}^{N} \ell(w_c, y_i, h_i) \tag{2}$$

Where $\ell(\bullet)$ is the loss function. We choose to use the quadratic hinge loss as the loss function which has the form of:

$$\ell(w_c, y_i, h_i) = [\max(0, w_c^T h_i y_i - 1)]^2 \tag{3}$$

We adopt the LBFGS algorithm [17] to solve the optimization problem (2). After learning the classifier parameters, we are able to predict the image categories with Eq. (1).

3 Experiments

We conduct experiments on the Oxford Flower datasets [1, 2] and the Scene-15 dataset [18] to evaluate the performance of the proposed color exemplar classifier for fine-grained image classification method. We densely extract local features of 16×16 pixels with an overlap of 6 pixels. The RGB, HSV, C-invariant, and opponent color spaces are used, as [8] did. To combine the spatial information, the spatial pyramid matching (SPM) approach [18] is adopted with $2^l \times 2^l, l = 0,1,2$. The codebook size is set to 1,024 for each channel. Sparse coding with max pooling [17] is used to extract the BoW representation of images for each channel. The one-versus-all rule is used for multi-class classification. We use the average of per-class classification rates as the quantitative performance measurement method. This process is repeated for several times to get reliable results.

Table 1. Performance comparison on the Oxford Flower 17 dataset

Methods	Performance
Nilsback and Zisserman [1]	71.76 ± 1.76
Varma and Ray [19]	82.55 ± 0.34
LP-B [20]	85.40 ± 2.40
KMTJSRC-CG [21]	88.90 ± 2.30
Ours	$\mathbf{91.53 \pm 1.39}$

3.1 Oxford Flower 17 Dataset

The first flower dataset we consider is a smaller one with 17 (*buttercup, colts' foot, daffodil, daisy, dandelion, fritillary, iris, pansy, sunflower, windflower, snowdrop, lily valley, bluebell, crocus, tigerlily, tulip* and *cowslip*) different flower categories [1]. This dataset has 1,360 images with 80 images for each class. For fair comparison, we follow the same experimental setup as [1] did and use the same training/validation/test (40/20/20) split. Figure 4 shows some example images of the Oxford Flower 17 dataset.

We compare the proposed color exemplar classifier based image classification method with [1, 19-21]. Nilsback and Zisserman [1] tried to learn the vocabularies for color, shape and texture features respectively. Varma and Ray [19] chose the most discriminative features by optimization. A boosting procedure is used by Gehler and Nowozin [20] to combine multiple types of features while Yuan and Yan [21] combined multiple types of features using a joint sparse representation approach.

Table 1 gives the performance comparison results. We can see from Table 1 that the proposed color exemplar classifier based image classification method outperforms other methods. This demonstrates the effectiveness of the proposed method. By combining color information per color channel, we are able to use the color information more efficiently. The effectiveness of considering each color channels separately can be seen from Table 1. The proposed method outperforms KMTJSRC-CG [21] by about 2.7 percent while [21] treated different color channels jointly. Besides, the use of exemplar classifier based semantic representation helps to represent image with more semantically meaningful histogram which helps to alleviate the semantic gap.

3.2 Oxford Flower 102 Dataset

The Oxford Flower 102 dataset is an extension of the Oxford Flower 17 dataset, hence is more difficult to classify than the Flower 17 dataset. This dataset has 8,189 images of 102 classes. Each class has 40-250 images. For each class, we use 10 images for training, 10 images for validation and the rest for testing, as [2] did for fair comparison. Figure 5 shows some example images of the Oxford Flower 102 dataset.

Fig. 5. Example images of the Oxford Flower 102 dataset

Table 2. Performance comparison on the Oxford Flower 102 dataset

Methods	Performance
Nilsback and Zisserman [1]	72.8
KMTJSRC-CG [21]	74.1
Ours	**76.3**

We compare the proposed color exemplar classifier based image classification method with [1, 21]. Table 2 gives the quantitative comparison results. We can have similar conclusions as on the Oxford Flower 17 dataset. The proposed color exemplar classifier based method outperforms the baseline methods [1, 21]. This again shows the effectiveness of the proposed method. As the Flower 102 dataset has more flower classes and large inter-class variation, a well chosen image representation is vital for the final image classification. This problem can be solved by using the proposed color exemplar classifier based representation hence helps to improve the classification performance.

3.3 Scene-15 Dataset

The last dataset we consider is the Scene-15 dataset [18]. This dataset consists of 15 classes of images (*bedroom, suburb, industrial, kitchen, livingroom, coast, forest, highway, insidecity, mountain, opencountry, street, tallbuilding, office* and *store*). Figure 6 shows some example images of this dataset. Each class has different sizes ranging from 200 to 400 images with an average of 300×250 pixel size. For fair comparison, we follow the same experimental procedure as [18] and randomly choose 100 images per class for classifier training and use the rest of images for performance evaluation.

We give the performance comparison of the proposed method with [17, 18, 22, 23, 24] in Table 3. We can see from Table 3 that the proposed method outperforms the baseline methods. Compared with exemplar based method [24], the use of color information can further improve the semantic representativeness of the exemplar based methods. These results demonstrate the proposed method's effectiveness.

Fig. 6. Example images of the Scene 15 dataset

Table 3. Performance comparison on the Scene 15 dataset

Methods	Performance
SPM [18]	81.40 ± 0.50
SPC [23]	81.14 ± 0.46
ScSPM [17]	80.28 ± 0.93
KC [22]	76.67 ± 0.39
WSR-EC [24]	81.54 ± 0.59
Ours	$\mathbf{83.75 \pm 0.52}$

4 Conclusion

This paper proposes a novel fine-grained image classification method using color exemplar classifiers. To combine the color information, we decompose each image into different color channels. We also train exemplar SVM classifiers for each channel and use the output of exemplar classifiers as the image representation for this channel to take advantage of the effectiveness of semantic based image representation. The final image representation is obtained by concatenating the representation of different channels. Experimental results on the Oxford Flower 17 and 102 datasets and the Scene-15 dataset demonstrate the effectiveness of the proposed method.

In our future work, we will consider how to encode the local features with smooth constraints [25, 26].

Acknowledgement. This work is supported in part by National Basic Research Program of China (973 Program): 2012CB316400; National Natural Science Foundation of China: 61303154, 61025011, 61272329, 61202325; the President Fund of UCAS; the Open Project Program of the National Laboratory of Pattern Recognition (NLPR); China Postdoctoral Science Foundation: 2012M520434, 2013T60156, 2013M530350. We thank Weigang Zhang and Qi Tian for their helpful suggestions.

References

1. Nilsback, M., Zisserman, A.: A visual vocabulary for flower classification. In: Proc. CVPR, pp. 1447–1454 (2006)
2. Nilsback, M., Zisserman, A.: Automated flower classification over a large number of classes. In: Proc. CVPR, pp. 722–729 (2008)
3. Yao, B., Bradski, G., Fei-Fei, L.: A codebook-free and annotation-free approach for fine-grained image categorization. In: Proc. CVPR, pp. 3466–3473 (2012)
4. Rosch, E., Mervis, C., Gray, W., Johnson, D., BoyesBraem, P.: Basic objects in natural categories. Cognitive Sci. 8(3), 382–439 (1976)
5. van de Weijer, J., Gevers, T., Bagdanov, A.: Boosting color saliency in image feature detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(1), 150–156 (2006)
6. Mindru, F., Tuytelaars, T., Gool, L., Moons, T.: Moment invariants for recognition under changing viewpoint and illumination. Computer Vision and Image Understanding 94(1-3), 3–27 (2004)
7. Bosch, A., Zisserman, A., Muoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(4), 712–727 (2008)
8. Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
9. Rasiwasia, N., Vasconcelos, N.: Holistic context models for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(5), 902–917 (2012)
10. Carneiro, G., Chan, A., Morena, P., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3) (2007)
11. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: Proc. CVPR, pp. 1577–1584 (2011)
12. Li, L., Su, H., Xing, E., Fei-Fei, L.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Proc. ECCV (2010)
13. Malisiewicz, T., Gupta, A., Efros, A.: Ensemble of exemplar-SVMs for object detection and beyond. In: Proc. ICCV (2011)
14. Zhang, C., Liu, J., Liang, C., Tang, J., Lu, H.: Beyond local image features: Scene classification using supervised semantic representation. In: Proc. ICIP (2012)
15. Jiang, W., Chang, S.-F., Loui, A.: Context-based concept fusion with boosted conditional random fields. In: Proc. ICASSP (2007)
16. Yang, Y., Huang, Z., Yang, Y., Liu, J., Shen, H., Luo, J.: Local image tagging via graph regularized joint group sparsity. Pattern Recognition 46(5), 1358–1368 (2013)
17. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proc. CVPR (2009)
18. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR (2006)
19. Varma, M., Ray, D.: Learning the discriminative power invariance trade-off. In: Proc. ICCV (2007)
20. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: Proc. ICCV (2009)
21. Yuan, X., Yan, S.: Visual classification with multi-task joint sparse representation. In: Proc. CVPR (2010)

22. Gemert, J., Veenman, C., Smeulders, A., Geusebroek, J.: Visual word ambiguity. IEEE Trans. Pattern Anal Machine Intell. 32(7), 1271–1283 (2010)
23. Zhang, C., Wang, S., Huang, Q., Liu, J., Liang, C., Tian, Q.: Image classification using spatial pyramid robust sparse coding. Pattern Recognition Letters 34(9), 1046–1052 (2013)
24. Zhang, C., Liu, J., Tian, Q., Liang, C., Huang, Q.: Beyond visual features: A weak semantic image representation using exemplar classifier for classification. Neurocomputing (2012), doi:10.1016/j.neucom.2012.07.056
25. Gao, S., Tsang, I., Chia, L.: Lappacian sparse coding, hypergraph laplacian sparse coding, and applications 35(1), 92–104 (2013)
26. Zhang, C., Wang, S., Huang, Q., Liang, C., Liu, J., Tian, Q.: Laplacian affine sparse coding with tilt and orientation consistency for image classification. Journal of Visual Communication and Image Representation 24(7), 786–793 (2013)