One Step Beyond Bags of Features: Visual Categorization Using Components

Jing Liu[1], Chunjie Zhang[1,3], Qi Tian[2], Changsheng Xu[1,3],Hanqing Lu[1], Songde Ma[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
{cjzhang, jliu, csxu, luhq}@nlpr.ia.ac.cn, masd@most.cn
[2]University of Texas at San Antonio, One UTSA Circle, San Antonio Texas, 78249-USA
qitian@cs.utsa.edu
[3]China-Singapore Institute of Digital Media

ABSTRACT

The bag-of-visual-words (BoW) representation has received wide application and public acceptance for visual categorization. However, the histogram based image representation ignores the spatial information and correlations among visual words. To tackle these problems, in this paper, we propose to use some image regions called 'components', as the higher-level visual elements to represent an image associating with the lower-level elements of 'visual words'. Then we formulate the task of visual categorization into two progressive relationships among a given concept and the two-level visual elements of images, i.e., visual-words-to-components and components-to-concept. Firstly, component level linear SVM classifiers are learned to model the relationship between visual words and components, then the output of these SVM classifiers are linearly combined to model the relationships between components and concept. Experiments on the Scene-15 dataset and the Oxford Flowers dataset demonstrate the effectiveness of the proposed method.

Index Terms— Visual categorization, component

1. INTRODUCTION

Automatic visual categorization is an important prerequisite to help people organize and access the increasing amounts of diverse multimedia data. Recent approaches to this problem rely more and more on the Bag-of-visual-Words (BoW) representation and the corresponding learning model, as they have shown promising performances in various tasks including image/object categorization [1-7]. Nonetheless, due to the loss of spatial information and correlation of local features, the discriminative power of the histogram based representation is limited. Considering large variations between images of the same class, how to extract representative structural descriptors and to further build a discriminative object model becomes a timely topic to address.

A lot of work has been made to improve the performance by utilizing the spatial layout information and correlations of visual words. Approaches using geometric correspondence search [1] achieved robustness at very high computational cost. Sivic et al. [2] used a more efficient approach by augmenting the basic bag-of-features representation with pairwise relation between neighboring local features. Grauman and Darrell [3] argumented features with their spatial coordinates in the pyramid matching kernel while Lazebnik et al. [4] proposed spatial pyramid matching method for natural scene recognition. These approaches employ some heuristic methods to boost the performance of visual categorization and may fail to provide the maximal discriminativity. To solve this problem, more and more researchers [5, 6] utilized learning methods to help make categorization of images. Li et al. [5] utilized multiple segmentations and then viewed object recognition as ranking holistic figure-ground hypotheses. Cao et al. [6] used heterogeneous features machines to learn the optimal combination of different types of features for visual recognition. Visual words were bundled together for large-scale near-duplicate image retrieval by Wu et al. [7] and the results were encouraging.

In this paper, we introduce 'component' – a set of image regions, as a higher-level element to represent an image jointly with the lower-level visual words, and propose a novel visual categorization model to build the correspondence among an image concept and the two-level visual elements, i.e., visual-words-to-components and components-to-concept. Firstly, to model the relationship between visual words and components, we try to learn component level linear SVM classifiers by assuming that each component has the same label as the image from which it is extracted. Since this assumption is weak and is probably contaminated with noise, instead of heuristically assign the weights, we then try to model the relationships between components and concept by learning a linear combination of the output of these SVM classifiers. This is achieved by minimizing the summed exponential loss between the predicted labels and the ground truth of training samples. Figure 1 shows an illustration of the proposed method for visual categorization using components.

The rest of the paper is organized as follows. Section 2 shows the proposed component based image representation. The details of making visual categorization using components are described in Section 3. Section 4 shows the

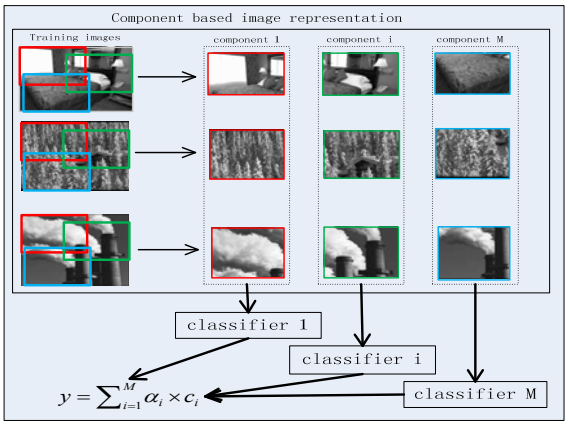

Figure 1: Illustration of the proposed method for visual categorization using components.

experimental results on both the Scene-15 dataset and the Oxford flowers dataset. The conclusion and future work are given in Section 5.

2. COMPONENT BASED IMAGE REPRESENTATION

In this paper, we propose to introduce 'components' as higher-level visual elements to represent an image associating with the lower-level elements of 'visual words'. A component is a set of image regions which can be generated using various methods, such as sampling, segmentation or detection. For each image, we densely extract its components with overlapping. The number of overlapped pixels varies depending on the size of each image to make sure the sampled components cover the whole image. For each component, we use the frequency distribution of visual words within each component as its feature representation. Hence, a component is higher-level representation and more descriptive than single visual word because it combines the spatial correlations among nearby visual words. We use the histogram based representation because it is invariant to rotation and efficient to compute. Note that other more descriptive representation methods (such as graph) can also be applied.

We will introduce the visual representations based on the two-level elements as follows. Formally, let $x_j^n \in \mathbb{R}^D$ be the j-th component of the n-th image, where D is the visual vocabulary size. The k-th element x_{jk}^n of x_j^n is the number of occurrence of visual word k within the j-th component of the n-th image. Components are arranged according to their relative positions within each image, as done in [8]. We use a matrix $\mathbf{X}^n = [\mathbf{x}_1^n, \mathbf{x}_2^n, \dots, \mathbf{x}_M^n] \in \mathbb{R}^{D \times M}$ to denote the n-th image, where M is the number of components. Note that if we regard the whole image as one component, this model will degenerate to the standard BoW representation. From this view, the BoW model can be viewed as a special case of our model.

Algorithm 1. Procedure of the optimization algorithm

1. **Input**: stopping threshold ε, training image set with labels $\{(\boldsymbol{c}^1, y^1), (\boldsymbol{c}^2, y^2), \dots, (\boldsymbol{c}^N, y^N)\}$, $index = 0$, $\boldsymbol{\alpha}_0$, $maxindex$.
2. While the reduced loss exceeds ε or $index < maxindex$
   $index = index + 1$;
   Calculate the gradient of $\boldsymbol{\alpha}_{index-1}$ using Eq. 5.
   Update $\boldsymbol{\alpha}_{index-1}$ to $\boldsymbol{\alpha}_{index}$ by gradient descent.
   Calculate the reduced loss of (3).
3. **Output**: The learned parameters of $\boldsymbol{\alpha}$

3. VISUAL CATEGORIZATION USING COMPONENTS

In this section, we will present the proposed model for visual categorization using components by modeling two progressive relationships among the two-level visual elements of images and a given concept, i.e., visual-words-to-components and components-to-concept.

3.1. Visual-Words-to-Components

To take the advantage of the component based image representation for visual categorization, a novel visual categorization method using components is proposed. For each image, we assume that each component within this image has a confidence value of the image category. However, we only have the label of images and it is often very hard to predict the label of components within each image. To alleviate this problem, we make some simplifications and assume that each component has the same label as the image from which it is extracted. We can then model the visual-word-to-component relationship by learning a mapping function f(x) from the training set.

Formally, Suppose we have a training image set with labels $\{(\boldsymbol{X}^1, y^1), (\boldsymbol{X}^2, y^2), \dots, (\boldsymbol{X}^N, y^N)\}$, where N is the number of training images and $y^n \in (-1,1)$ is the label of the n-th image, $n \in \{1,2,\dots,N\}$. Using the same symbols as above, we first split the training set by component into M sets as $\{(\boldsymbol{x}_m^1, y^1), (\boldsymbol{x}_m^2, y^2), \dots, (\boldsymbol{x}_m^N, y^N)\}$ with $m \in \{1,2,\dots,M\}$ is the indices of components and M is the number of components within each image, as is shown in figure 1. This is based on the observation that: the relatively same components within images of the same class often have similar visual representations; this positional information can help make correct categorization of images. For example, images of beach often have sky on the upper side while sand on the lower side; however, images of street often have sky on the upper side and buildings on the left and right sides of images.

After splitting the training dataset by components, M linear SVM classifiers can then be trained to model the relationships between visual words and components. We choose the linear SVM classifier because it is efficient to compute and is also robust to noise. Then we can use the learned classifiers to predict the confidence values of each component as $f_m(x_m^1), m \in \{1,2,\dots,M\}$.

3.2. Components-to-Concept

After the visual-words-to-component level relationship is learned, we can then try to predict the categories of images. We choose to linearly combine the component level results to predict the category of images as follows:

$$\widehat{y^n} = \sum_{m=1}^{M} \alpha_m \times f_m(x_m^n) \quad (1)$$

Where $\widehat{y^n}$ is the predicted label for $n-th$ image, let $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_M]^T$ and $\boldsymbol{c^n} = [f_1(x_1^n), f_2(x_2^n), \dots, f_M(x_M^n)]^T$, we can rewrite Eq.1 as:

$$\widehat{y^n} = \boldsymbol{\alpha}^T \times \boldsymbol{c^n} \quad (2)$$

The parameter of $\boldsymbol{\alpha}$ can be learned using the training set by solving the optimization problem as:

$$[\boldsymbol{\alpha}] = \mathrm{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{L}(\mathrm{y}^{\mathrm{n}}, \boldsymbol{\alpha}^{\mathrm{T}} \times \mathbf{c}^{\mathrm{n}}) \quad (3)$$

where $\mathrm{L}(\cdot,\cdot)$ is the loss function. It penalizes the discrepancy between the predicted label and the ground truth. The loss function can be the least squares loss, the hinge loss or the exponential loss. It has been shown that, the hinge loss and exponential loss are more efficient than the least squares loss in classification. Here we choose the exponential loss because it is differentiable and efficient to implement. The exponential loss has the following form as:

$$\mathrm{L}(\mathrm{y}, \boldsymbol{\alpha}^{\mathrm{T}} \times \mathbf{c}) = \exp(-\mathrm{y} \times \boldsymbol{\alpha}^{\mathrm{T}} \times \mathbf{c}) \quad (4)$$

Problem (3) is convex in $\boldsymbol{\alpha}$, hence the global optimal can be found using gradient descent as:

$$\frac{\partial L(y, \boldsymbol{\alpha}^T \times \boldsymbol{c})}{\partial \boldsymbol{\alpha}} = \qquad = -y \times \boldsymbol{c} \times exp\,(-y \times \boldsymbol{\alpha}^T \times \boldsymbol{c}) \quad (5)$$

After the parameters are learned, we can predict the category of images using Eq. 2. Algorithm 1 shows the procedure of learning the parameters of $\boldsymbol{\alpha}$ by solving problem (3). It can be stopped either the reduced loss is below a threshold or the iteration step exceeds a predefined number of steps.

4. EXPERIMENTS

We evaluate the proposed method to category and scene classification tasks on two public datasets: Scene-15 dataset from Lazebnik *et al* [4] and Oxford Flowers dataset from Nilsback and Zisserman [9]. The Scene-15 dataset composes 4,485 images, which range from man-made environments like offices and kitchens to natural scenes like forests and mountains. The Oxford Flowers dataset contains 17 categories of flowers with 80 images per category. As to feature extraction, we follow the same parameter setting as did in [4, 10] and densely compute SIFT descriptors on overlapping 16×16 pixels with an overlap of 8 pixels. We extract 5×5 components with overlapping for each image. The number of overlapped pixels varies depending on the size of images to make sure the sampled components cover the whole image. Multi-class classification is done via the one-versus-all rule: a classifier is learned to separate each class from the rest and a test image is assigned the label of the classifier with the highest response. The classification performance is measured quantitatively by the average of per-class classification rates.

Methods	Classification rate
Lazebnik *et al*. [4]	81.4%
Gemert *et al*. [10]	76.3%
Ours	**84.1%**

Table 1. Performance comparison of different methods and the proposed method on the Scene-15 dataset (100 training images per class). The best result is in bold.

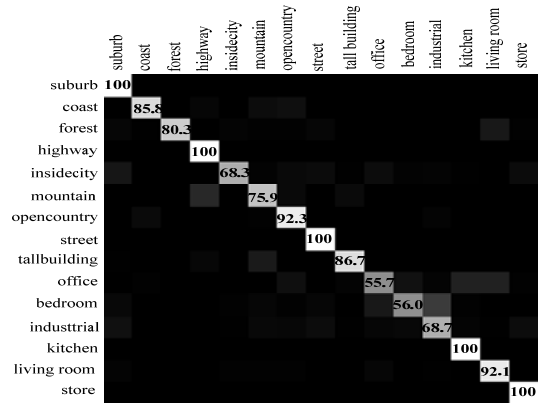

Table 2. Confusion table for the scene category dataset. Classification rates for individual classes are listed along the diagonal.

4.1. Scene-15 Dataset

The first dataset we consider is the Scene-15 dataset. The major pictures include the COREL collection, Google Image Search and personal photographs. Each category has 200 to 400 images with sizes of about 300×250 pixels. We use the same number of training images per category as [4, 10] and choose the first 100 images per category as the training set and use the remaining images as the test set. A codebook is created by *k*-means clustering with 1,000 clusters.

Table 1 shows the performance comparison on the Scene-15 dataset for the proposed method and methods in [4] and [10]. Lazebnik *et al*. [4] utilized spatial pyramid matching along with SVM classifiers while Gemert *et al*. [10] made image classification by modeling soft-assignment in the popular codebook model. Our classification rate is 84.1%, which is better than the results of [4] and [10]. This demonstrates the effectiveness of the proposed method for modeling the relationship between visual features and image concept. Compared with histogram based image representation, our component based image representation can preserve more spatial relationship and correlations of visual words; besides, instead of heuristically combine the category information of different components, we efficiently and effectively learn these weights by utilizing an

Methods	Classification rate
Nilsback and Zisserman[9]	71.76±1.76
Varma and Ray [12]	82.55±0.34
χ^2 [13]	87.45±1.13
Ours	**88.24±1.28**

Table 3. Performance comparison of different methods and the proposed method on the Oxford Flowers dataset. The best result is in bold.

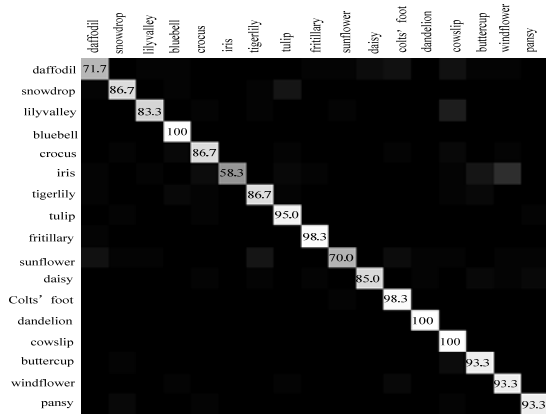

Table 4. Confusion table for the Oxford Flowers dataset. Classification rates for individual classes are listed along the diagonal.

optimization procedure. To analysis the class level results, Table 2 also gives the confusion matrix for the Scene-15 dataset. We can see from Table 2 that the indoor classes are more difficult than the outdoor classes.

4.2. Oxford Flowers dataset

The second dataset we consider is the Oxford Flowers dataset. For each category, 40, 20 and 20 images are used for training, validation and test respectively. For fair comparison, we use the three splits provided by the authors of [9]. Since color information is also very important for flower recognition, besides SIFT features, we also extract C-SIFT features [11] which has been shown very effective to improve classification performances. A codebook with 1,000 clusters is created by *k*-means clustering for the SIFT and C-SIFT feature respectively. We combine the results by taking the averages of the predicted values of using SIFT and C-SIFT features.

Table 3 shows the performance comparison on the Scene-15 dataset for the proposed method and methods in [9, 12 and 13]. Nilsback and Zisserman [9] tried to distinguish one flower from another by developing a visual vocabulary that explicitly represents the various aspects (color, shape and texture). Varma and Ray tried to learn the most discriminative features for categorization. Xie et al [13] used χ^2 distances to measure the dissimilarity. We also give the confusion matrix for the Oxford Flowers dataset in Table 4. This is achieved by taking the average of the three data-split results. We can have similar observation as on the Scene-15 dataset. The results on both the Scene-15 dataset and the Oxford Flowers dataset demonstrate the effectiveness of the proposed method.

5. CONCLUSION

This paper focuses on the task of object recognition with an enhanced image representation. Generally, two keypoints or contributions are addressed in our work. (1) A novel component based image representation is proposed to model the spatial information and correlations of visual words. (2) We formulate the task of visual categorization into two progressive relationships among a given concept and the two-level visual elements of images, i.e., visual-words-to-components (SVM classifier) and components-to-concept (linear combination). The comprehensive experimental evaluations on the two public datasets of Scene-15 dataset and the Oxford flowers dataset demonstrate the effectiveness of the proposed method.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 60903146, 60835002 and 90920303) and 973 Program (Project No. 2010CB327905).

6. REFERENCES

[1] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proc. CVPR*, volume 1, pages 26-33, 2005.

[2] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *Proc. ICCV*, 2005.

[3] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Proc. ICCV*, pp.1458-1465, 2005.

[4] S. Lazebnik, C. Schmid, J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006.

[5] F. Li, J. Carreira and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *Proc. CVPR*, 2010.

[6] L.Cao, J. Luo, F. Liang and T. Huang. Heterogeneous features machines for visual recognition. In *Proc. ICCV*, 2009.

[7] Z. Wu. Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *Proc. CVPR*, 2009.

[8] X. Liu, B. Cheng, S. Yan, J. Tang, T. Chua, and H. Jin. Label to region by bi-layer sparsity priors. In *ACM MM*, 2009.

[9] M. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Proc. CVPR*, 2006.

[10] J. Gemert, C. Veenman, A. Smeulders and J. Geusebroek. Visual word ambiguity. In IEEE *Transactions and Pattern Analysis and Machine Intelligence*. 2010.

[11] K. Sande, T. Gevers and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010.

[12] M. Varma and D. Ray. Learning the discriminative power invariance trade-off. In *Proc. ICCV*, 2007.

[13] N. Xie. H. Ling, W. Hu and X. Zhang. Use bin-ratio information for category and scene classification. In *Proc. CVPR*, 2010.