Soft Measure of Visual Token Occurrences for Object Categorization

Yanjie Wang, Xiabi Liu*, and Yunde Jia

Beijing Laboratory of Intelligent Information Technology, School of Computer
Science, Beijing Institute of Technology
{wangyanjie,liuxiabi,jiayunde}@bit.edu.cn
Tel.: +86-10-68913447, Fax: +86-10-86343158

Abstract. The improvement of bag-of-features image representation by statistical modeling of visual tokens has recently gained attention in the field of object categorization. This paper proposes a soft bag-of-features image representation based on Gaussian Mixture Modeling (GMM) of visual tokens for object categorization. The distribution of local features from each visual token is assumed as the GMM and learned from the training data by the Expectation-Maximization algorithm with a model selection method based on the Minimum Description Length. Consequently, we can employ Bayesian formula to compute posterior probabilities of being visual tokens for local features. According to these probabilities, three schemes of image representation are defined and compared for object categorization under a new discriminative learning framework of Bayesian classifiers, the Max-Min posterior Pseudo-probabilities (MMP). We evaluate the effectiveness of the proposed object categorization approach on the Caltech-4 database and car side images from the University of Illinois. The experimental results with comparisons to those reported in other related work show that our approach is promising.

1 Introduction

In recent years, object categorization with bag-of-features image representation has become a hot topic in the field of compute vision and pattern recognition [1,2,3,4,5,6,7,8]. The bag-of-features method originated from the bag-of-words model for document analysis, which was firstly introduced to object categorization by Csurka et al. [1]. They cluster local features in images by k-means algorithm to generate a visual vocabulary. The image is then represented as a histogram over visual tokens in the vocabulary. After Csurka et al.'s work, statistical modeling of visual tokens has been advised to improve the effectiveness of bag-of-features representation [2,3,6]. The relations between local features and visual tokens can be described more accurately and reliably through statistical modeling of visual tokens. Furthermore, a local feature is allowed to be softly mapped to multiple visual tokens in this way, so the aliasing effects can be reduced. In existing

* The corresponding author.

X. Jiang and N. Petkov (Eds.): CAIP 2009, LNCS 5702, pp. 774–782, 2009.

methods of statistical modeling of visual tokens for object categorization, the Gaussian distribution is used to model each visual token and the set of visual tokens is considered as a Guassian Mixture Model (GMM) [2,3,6].

This paper proposes a new soft bag-of-features image representation based on the Gaussian Mixture Modeling (GMM) of visual tokens. The resultant object categorization approach includes four stages. Firstly, local features are extracted from an input image. Secondly, posterior probabilities of being visual tokens for local features are computed by using Bayes formula, where local features from each visual token are assumed to be of the distribution of GMM. The GMM is learned from the training data by the Expectation-Maximization (EM) algorithm with a model selection method based on the Minimum Description Length (MDL). Thirdly, the image is represented using one of three schemes: probabilities based hard histogram, classification based soft histogram, and completely soft histogram. Finally, the image is classified into one of object categories under a new discriminative learning framework of Bayesian classifiers, the Max-Min posterior Pseudo-probabilities (MMP) [9], where feature vectors of images from each object category is also assumed to be of the distribution of GMM. Following other related work, we evaluate the proposed object categorization approach on the Caltech-4 database and the car side images of the University of Illinois. Our approach experimentally outperforms some other related methods with the similar local features and achieves the comparable results to those reported by using more sophisticated local features.

2 GMM-MMP Classification Framework

In this section, we introduce posterior pseudo-probabilities based categorization approach with the MMP learning. The reader is referred to our paper for more details [9].

Let $\mathbf{X}$ be a feature vector, C be an object category, $p(\mathbf{X}|C)$ be the class-conditional probability density function, then the posterior pseudo-probability of being C for $\mathbf{X}$ is computed as

$$f(p(\mathbf{X}|C)) = 1 - \exp(-\lambda p^{\theta}(\mathbf{X}|C)), \tag{1}$$

where λ, θ are positive numbers. Consequently, $f(p(\mathbf{X}|C))$ is a smooth, monotonically increasing function of $p(\mathbf{X}|C)$, and $f(0) = 0$ and $f(+\infty)$. Given an input image, we use Eq. 1 to compute the posterior pseudo-probability for each object category. The category with maximum posterior pseudo-probability will be taken as the categorization result.

The MMP method is advised to learn unknown parameters in Eq. 1. Let $f(\mathbf{X};\mathbf{\Lambda})$ be the posterior pseudo-probability measure function (Eq. 1) of an object category, where Λ denote the set of unknown parameters in it. Let $\hat{\mathbf{X}}_i$ be a feature vector of arbitrary positive sample of the category, $\bar{\mathbf{X}}_i$ be a feature vector of arbitrary negative sample of the category, m and n be the number of positive and negative samples, respectively. Then the objective function of MMP learning for estimating parameters is

$$F(\mathbf{\Lambda}) = \frac{1}{m}\sum_{i=1}^{m}[f(\hat{\mathbf{X}}_i;\mathbf{\Lambda}) - 1]^2 + \frac{1}{n}\sum_{i=1}^{n}[f(\bar{\mathbf{X}}_i;\mathbf{\Lambda})]^2. \tag{2}$$

$F(\mathbf{\Lambda}) = 0$ means the perfect classification performance on the training data. Consequently, we can obtain the optimum parameter set $\mathbf{\Lambda}^*$ of the posterior pseudo-probability measure function by using the gradient descent algorithm to minimize $F(\mathbf{\Lambda})$:

$$\mathbf{\Lambda}^* = \arg\min_{\mathbf{\Lambda}} F(\mathbf{\Lambda}). \tag{3}$$

The form of class-conditional probability density function $p(\mathbf{X}|C)$ in Eq. 1 should be provided for using MMP categorization framework, which is assumed to be the GMM in this paper. Let K be the component number of GMM, w_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ be the weight, the mean, and the covariance matrix of the k-th Gaussian component, respectively. $\sum_{k=1}^{K} w_k = 1$. Then we have

$$p(\mathbf{X}|C) = \sum_{k=1}^{K} w_k N(\mathbf{X}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{4}$$

where

$$N(\mathbf{X}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{X} - \boldsymbol{\mu}_k)\right). \tag{5}$$

So the set of unknown parameters in the posterior pseudo-probability measure function (Eq. 1) of each object category is

$$\mathbf{\Lambda} = \{\lambda, \theta, w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, k = 1, \cdots, K. \tag{6}$$

3 Object Categorization by Soft Measure of Visual Token Occurrences

In bag-of-features image representation, a visual vocabulary consisting of visual tokens is generated to bridge local features and images. In this paper, we model the distribution of local features from each visual token as a GMM. The corresponding visual vocabulary can be seen as a set of visual token GMMs. According to visual token GMMs, we compute posterior probabilities of being visual tokens for local features. Then three corresponding image representation schemes are explored for object categorization under the GMM-MMP categorization framework.

3.1 Visual Token GMM with MDL-EM Training

We firstly cluster local features extracted from training images into designated number of groups. Each group of local features is corresponding with a visual token. This is the same as conventional bag-of-features methods. However, each group of local features is represented by a GMM, instead of its center, in this paper. The GMM is fitted to the group data by using the Expectation-Maximization (EM) algorithm [10] with a model selection method based on the Minimum Description Length (MDL) [11]. This strategy makes our method different from other statistical modeling based bag-of-features representations, where GMM is used to model the whole vocabulary and each visual token is corresponding with a Gaussian component.

Let t be the parameter number of each Gaussian component in the GMM, n be the number of training samples, $f(\mathbf{x}_1, \cdots, \mathbf{x}_n|\Theta)$ be the likelihood function over the training set. Then the training criterion with MDL-EM can be formalized to minimize [11]

$$-\log f(\mathbf{x}_1, \cdots, \mathbf{x}_n|\Theta) + \frac{t}{2}\log n, \tag{7}$$

where the first and second terms stand for the objective of maximum likelihood and the simplest model, respectively.

After visual token GMMs are obtained from the training data, we employ Bayes formula with the assumption of the same prior probabilities for all the visual tokens to estimate the posterior probability of being visual token $\mathbf{v}_j$ for local feature $\mathbf{x}_i$:

$$P(\mathbf{v}_j|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|\mathbf{v}_j)}{\sum_{k=1}^{N} P(\mathbf{x}_i|\mathbf{v}_k)}, \tag{8}$$

where N is the number of all the visual tokens.

3.2 Image Representation

According to hard assignment of local features to visual tokens, it seems that we can only compute occurrence frequencies of visual tokens to obtain a hard histogram description of the image. Oppositely, $P(\mathbf{v}_j|\mathbf{x}_i)$ reflects the confidence of assigning $\mathbf{x}_i$ to $\mathbf{v}_j$. More reliable and accurate occurrence distribution can be defined based on this soft assignment. In this paper, we consider three corresponding representation schemes: Probability Based Hard Histogram (PBHH), Classification Based Soft Histogram (CBSH), and Completely Soft Histogram (CSH). In both PBHH and CBSH, local features are firstly classified into the visual token with maximum posterior probability. Then the image is represented as frequencies (PBHH) or mean probabilities (CBSH) of visual tokens. The CSH maps each local feature to all the visual tokens, and compute mean probabilities of visual tokens to represent the image. More formally, let I be an image, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ be local features extracted from I, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$ be tokens in the visual vocabulary, M be the number of all the local features, $m_i|_{i=1}^{N}$ be the number of local features classified into visual token $\mathbf{v}_j$, then these three representation schemes are listed in Table 1.

After the image is represented by each of three schemes above, we assume that the feature vectors of the images from each object category are of the distribution of GMM. We then use GMM-MMP categorization framework descried in Section 2 to perform the image categorization.

Table 1. Three image representation schemes

Schema	Image Representation
PBSH	$\{m_i/M\}\|_{i=1}^N$
CBSH	$\left\{\sum_{k=1}^{m_i} P(\mathbf{v}_i\|\mathbf{x}_k)/m_i\right\}\|_{i=1}^N$
CSH	$\{\sum_{k=1}^{M} P(\mathbf{v}_i\|\mathbf{x}_k)/M\}\|_{i=1}^N$

4 Experimental Results

4.1 Experimental Setup

Harris-affine detector [12] is adopted to extract local features from images. Then we use SIFT [13] as the feature descriptor, resulting a 128-dimentsional real vector ($4 \times 4 \times 8$) for each local feature. In order to fairly compare our approach to other related methods, the number of visual tokens is preset to 1000 as that used in [1].

In the MMP training for object categories, positive samples of each object category are images from this category, while negative ones are images from other categories. At first, we used the MDL-EM algorithm on positive samples to get the parameters in the GMM, and set λ and θ through experiments. Then the MMP training algorithm was used on all the samples including positive samples and negative samples to revise initial parameters obtained by the MDL-EM algorithm. For the MDL based model selection of the GMM, we evaluate the component numbers from 1 to 20 for visual tokens and object categories. The resultant numbers of components for visual tokens vary from 1 to 5, while those for object categorization are from 3 to 9.

4.2 Caltech-4 Database and Car Side Images

We conduct experiments of object categorization on the Caltech-4 database and car side images from the University of Illinois. The Caltech-4 database includes four object categories. The number of images from each category varies from 450 to 1074. Following the evaluation method used in other related work, we randomly select half of images from each object category for training, and the others for testing.

4.2.1 Comparative Evaluation

The proposed object categorization approach can be divided into three stages: visual token modeling, image representation, and discriminative learning for object categorization. Thus we design three groups of experiments to evaluate the influence of various factors in each stage.

In the first group, we tested the effectiveness of three visual token modeling methods including GMM, Gaussian Model, and traditional cluster center under hard histogram image representation setting. Gaussian Model is treated as 1-component GMM in our experiments. We also compared GMM and Gaussian Model under two types of soft histogram image representation setting, CBSH

Table 2. Comparing categorization accuracies for visual token modeling methods and image representation schemes

Methods	THH	PBHH		CBSH		CSH	
		GM	GMM	GM	GMM	GM	GMM
Airplanes	0.961	**0.968**	**0.968**	0.972	**0.980**	**0.985**	0.970
Cars(Rear)	0.960	0.964	**0.974**	0.945	**0.971**	0.945	**0.978**
Motorbikes	0.889	0.889	**0.910**	0.893	**0.910**	0.893	**0.932**
Faces	0.880	0.893	**0.907**	0.880	**0.889**	0.906	**0.933**
Cars(Side)	0.953	**0.942**	0.935	0.956	**0.960**	0.956	**0.964**
Mean	0.936	0.939	**0.947**	0.936	**0.952**	0.943	**0.960**

Table 3. Comparing categorization accuracies for MMP vs. MDL-EM training

Categories	**MMP**	**EM**
Airplanes	**0.980**	0.970
Cars(Rear)	**0.995**	0.978
Motorbikes	**0.960**	0.932
Faces	**0.947**	0.933
Cars(Side)	**1.000**	0.964
Mean	**0.977**	0.960

and CSH. In this group, only MDL-EM algorithm is used to learn the GMMs of object categories. The MMP algorithm is not triggered yet. Table 2 shows categorization results for the test data from 5 categories, where 'THH' denotes the Traditional Hard Histogram based on cluster centers, 'GM' denotes Gaussian Model. It demonstrates that the GMM behaved best and statistical modeling of visual tokens brings better performance than distance based vector quantization technique. In the second group, three proposed image representation schemes, PBHH, CBSH, and CSH, are compared under the GMM of visual tokens. The corresponding results are also reflected in Table 2, where CSH is shown to outperform other two schemes. We tested the effectiveness of the MMP discriminative learning algorithm in the third group. Under CSH image representation with the GMM of visual tokens, the training effects of the MMP and the MDL-EM for object categorization were compared and listed in Table 3. It shows that the mean categorization accuracy is improved from 96.0% (EM) to 97.7% (MMP).

4.2.2 Comparisons to Related Work

To confirm the effectiveness of our approach, we further compared our best categorization results achieved by using the GMM of visual tokens, CSH image representation, and MMP learning algorithm to those reported in other related work [1, 4, 14, 7, 15]. The comparisons of results are shown in Table 4. Among these work under comparison, Csurka et al. [1] and Sivic et al. [14] adopt the same local features as ours, namely, the Harris-affine detector with the SIFT descriptor; Fergus et al. [7] uses the Kadir-Brady local feature detector and the pixel descriptor; Kapoor et al. [15] employs the multiresolution local features;

Table 4. The comparisons between our approach and other related methods

Categories	**Ours**	[1]	[14]	[7]	[4]-1	[4]-2	[15]
Airplanes	0.980	0.963	0.953	0.902	0.889	0.975	0.980
Cars(Rear)	0.995	0.977	0.981	0.900	0.911	1.000	0.991
Motorbikes	0.960	0.927	0.936	0.925	0.922	0.943	0.970
Faces	0.947	0.940	0.940	0.964	0.935	1.000	0.995
Cars(Side)	1.000	0.996	–	–	0.830	1.000	–
Mean	0.977	0.961	–	–	0.897	0.984	–

Opelt et al. [4] tested two kinds of local features, including the affine invariant interest point detector with the moment invariant descriptor (denoted as [4]-1 in Table 4) and the similarity-measure-segmentation with the intensity distribution description (denoted as [4]-2 in Table 4). Our approach experimentally outperforms the methods using the similar local features [1,14], [4]-1, and achieved the comparable results to those reported by using more sophisticated local features in [15] and [4]-2.

5 Conclusions

In this paper, we explored the problem of soft histogram image representation based on Gaussian Mixture Modeling (GMM) of visual tokens for object categorization. The main contributions of this paper are summarized as follows: 1) The posterior probabilities of being visual tokens for local features are computed by assuming that local features from each visual token are of the distribution of GMM. Accordingly, three types of image descriptions are defined and compared for object categorization, including Probability Based Hard Histogram (PBHH), Classification Based Soft Histogram (CBSH), and Completely Soft Histogram (CSH). 2) A new discriminative learning framework of Bayesian classifiers, Max-Min posterior Pseudo-probabilities (MMP), is applied to object categorization.

We conducted three groups of comparative experiments on the Caltech-4 database and car side images from the University of Illinois. In the first group, GMM of visual tokens is compared to Gaussian modeling of visual tokens as well as traditional cluster center. The results show that the GMM outperforms other two strategies. In the second group, three types of histogram descriptions of the images are tested and CSH is shown to behave best. In the last group, we demonstrate that MMP is better than generative learning counterpart. To sum up, we achieved the best result by using the GMM of visual tokens, CSH image representation, and MMP learning for object categorization, which is better than those reported using similar local features and comparable to those obtained from more sophisticated local features.

The future developments of the proposed approach are described as follows. Firstly, visual token GMMs are currently learned by using the MDL-EM algorithm, since the number of visual token is 1000 and the MMP is not enough efficient to solve this huge classification problem. In the next work, we will improve the efficiency of MMP learning and apply it to the training of visual token GMMs for more accurate measure of visual token occurrence. Secondly, the experimental evaluation of our approach is planned to be performed on other widely used databases, including Caltech-101 and VOC 2008.

Acknowledgement

This research was partially supported by 973 Program of China (No. 2006CB303103), Beijing Key Discipline Program, and Excellent Young Scholars Research Fund of Beijing Institute of Technology (No. 2008YS1203).

References

1. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision in ECCV (2004)
2. Farquhar, J., Szedmak, S., Meng, H., Shawe-Taylor, J.: Improving "bag-of-keypoints" image categorisation: Generative models and pdf-kernels. Technical report, University of Southampton (2005)
3. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: IEEE International Conference on Computer Vision (2005)
4. Opelt, A., Fussengger, M., Pinz, A., Auer, P.: Generic object recognition with boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(3), 416–513 (2006)
5. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)
6. Perronnin, F.: Univeral and adapted vocabularies for generic visual categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(7), 1243–1256 (2008)
7. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE conference on Computer Vision and Pattern Recognition (2003)
8. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE conference on Computer Vision and Pattern Recognition (2006)
9. Liu, X., Jia, Y., Chen, X., Deng, Y., Fu, H.: Image classification using the max-min posterior pseudo-probabilities method. Technical Report BIT-CS-20080001, Beijing Institute of Technology (2008), http://www.mcislab.org.cn/member/~xiabi/papers/2008_1.PDF
10. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society 39(1), 1–38 (1977)
11. Hansen, M.H., Yu, B.: Model selection and the principle of minimum description length. Journal of American Statistical Association 96(454), 746–774 (1989)
12. Mikolajczyk, K., Shmid, C.: Scale and affine invariant point detectors. International Journal of Computer Vision 60(1), 63–86 (2004)

13. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision (1999)
14. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: IEEE International Conference on Computer Vision (2005)
15. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with gaussian processes for object categorization. In: IEEE International Conference on Computer Vision (2007)