Pattern Recognition Letters 31 (2010) 702–708

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Novel Gaussianized vector representation for improved natural scene categorization

Xi Zhou *, Xiaodan Zhuang, Hao Tang, Mark Hasegawa-Johnson, Thomas S. Huang

Beckman Institute of Advanced Science and Technology, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Available online 16 December 2009

Keywords:
Gaussian mixture model
Unsupervised learning
Image classification
Expectation–Maximization

ABSTRACT

We present a novel Gaussianized vector representation for scene images by an unsupervised approach. Each image is first encoded as an ensemble of orderless bag of features. A global Gaussian Mixture Model (GMM) learned from all images is then used to randomly distribute each feature into one Gaussian component by a multinomial trial. The posteriors of the feature on all the Gaussian components serve as the parameters of the multinomial distribution. Finally, the normalized means of the features distributed in every Gaussian component are concatenated to form a supervector, which is a compact representation for each scene image. We prove that these supervectors observe the standard normal distribution. The Gaussianized vector representation is a more generalized form of the widely used histogram representation. Our experiments on scene categorization tasks using this vector representation show significantly improved performance compared with the histogram-of-features representation. This paper is an extended version of our work that won the IBM Best Student Paper Award at the 2008 International Conference on Pattern Recognition (ICPR 2008) (Zhou et al., 2008).

1. Introduction

Growing research attention has been put on analysis and recognition of natural scenes (forest, street, office, etc.). Most previous approaches to this problem leveraged on supervised segmentation as preprocessing or the manual annotation of "intermediate" properties (Treisman and Gelade, 1980; Oliva and Torralba, 2001; Vogel and Schiele, 2004). These properties might be considered as classes of texture information.

In recent years, bag-of-features methods, which represent an image as an orderless collection of local features, have demonstrated good performance (Wallraven et al., 2003; Willamowski et al., 2004; Grauman and Darrell, 2005; Fei-Fei and Perona, 2005) for the whole-image categorization tasks. Furthermore, Lazebnik proposed to adopt spatial pyramid matching for scene categorization in order to utilize the spatial information beyond the bag-of-features image representation (Lazebnik et al., 2006). In particular, all these work calculate the histogram-of-features as representation used for classification. Fei-Fei and Perona (2005) show that the histogram-of-features representation outperforms previous keyword matching approach for bag-of-features.

A common difficult that arises for the task of natural scene categorization and many other general tasks of computer vision is to find correspondences among multiple images. That is, how do we match the corresponding feature points between pairs of images? Many dimensionality reduction techniques, including the global linear transformation methods such as Principal Component Analysis (PCA) and Fisher's Linear Discriminant Analysis (LDA) as well as the manifold learning methods such as Locally Linear Embedding (LLE) and Locality Preserving Projections (LPP), require well-corresponded feature points between the images to seek a meaningful low dimensional subspace. For the various classifiers based on certain distance metrics in the feature space, such as the Nearest Neighbor (NN), correspondences are critical, too. For instance, it is meaningless to compute the distance between the nose tip point in one face image and the left eye corner point in another face image. The challenges of finding correspondences between images are at least two folds: First, in many cases the images undergo certain unknown transformations (e.g., rotation, affine, etc.) and the features extracted from these images are correspondingly distorted. Although such distortions can be somehow compensated by adopting features relatively invariant to the transformations, such as the Scale Invariant Feature Transform (SIFT) descriptor, it is unlikely to be possible to reverse the effect of the unknown transformations. Second, the order of the extracted feature vectors are partially, if not completely, unknown. This makes the task of finding two corresponding feature vectors in two different images extremely difficult.

Natural scene categorization turns out to present extra challenges for correspondences compared to other computer vision tasks in general. For example, in the tasks of facial information recognition, such as face recognition as well as age and head pose estimation, it is possible, although challenging, to obtain correspondences by performing global alignment of the face images (Wang et al., 2002; Jiao et al., 2003). Therefore, it is natural to

* Corresponding author. Fax: +1 217 333 2922.
E-mail address: xizhou2@uiuc.edu (X. Zhou).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.12.010

represent an aligned face image as a vector of features (ordered by their spatial locations) for the appearance based methods. The features for natural scene categorization, however, are more complicated. The scene images, even though they belong to the same scene category, have various spatial layout. For example, an entire bed or a partial bed with different styles can be seen from different directions (viewing angles) in different images all of which belong to the "bedroom" category, and different types of furniture like chairs and desks can be present in the images to add the ambiguity. Therefore, global alignment provides limited correspondences between natural scene images, because the inhomogeneous regions of the scene images are still misaligned. Image segmentation can help to find the correspondences to a certain degree. However, image segmentation itself is another hard problem, especially when the spatial structure of the image becomes more and more complicated. This motivates the need of a robust, efficient, and geometrically invariant structural scene image representation.

On the other hand, many dimensionality reduction algorithms, such as PCA and LDA, are based on the implicit assumption that the features observe the Gaussian distribution. In other words, the results of PCA and LDA are optimal only when the features are Gaussian distributed. Moreover, many classifiers are based on the Euclidean distance, which is a commonly used distance metric between two vectors. Therefore, it is desirable to design a vector-based image representation that observes the standard Normal distribution.

In the classical histogram-of-features representation Schiele and Crowley (2000), Swain and Ballard (1991), the histogram bins are chosen by a *k*-means algorithm on the whole patch data. Then each patch is distributed to a particular bin based on its distance to the cluster centroids. However, histogram representation has some intrinsic limitations. For example, it is sensitive to several factors such as outliers, the choice of bins, and the noise level in the data. Most importantly, encoding high-dimensional feature vectors by a relatively small codebook inclines to large quantization errors and lose of discriminability Boiman et al. (2008).

In this paper, we present a novel approach to transform the scene images into correspondent and normalized feature vectors in an unsupervised manner. First, each scene image is encoded as an ensemble of orderless bag of features. The features from all the images are used to train a global Gaussian Mixture Model (GMM). For every image, this global GMM with M Gaussian components is used to randomly distribute each feature of the image into one of the M classes (i.e., Gaussian components) by a multinomial trial. In particular, we calculate the posterior probabilities of the feature against all the M Gaussian components and use these posterior probabilities as the parameters for the multinomial trial. Finally, the normalized means of the features which are distributed into every class are concatenated to form a super-vector, which is a compact representation for the scene image. We justify that such feature vectors from our new representation observe the standard normal distribution. We demonstrate the effectiveness of this novel Gaussianized vector-based image representation through the natural scene categorization task on a 15 scene category database. Our experiment results show that significantly better performance is achieved by our feature representation than is achieved by the traditional histogram representation for bag-of-features. In addition, our method outperforms the system with probabilistic latent semantic analysis (pLSA) (Hofmann, 2001) and spatial pyramid matching (Lazebnik et al., 2006).

The rest of the paper is organized as follows: Section 2 introduces the Gaussianized vector representation for natural scene images, and shows that correspondences and Gaussianization are achieved in the proposed representation. Section 3 connects our proposed representation to the widely used histogram-of-features representation and points out that the histogram-of-features representation can be viewed as a special case of our representation. Section 4 describes the patch representations that are adopted in this work. Section 5 presents our experiment results on both the 13-class and 15-class natural scene categorization tasks. Section 6 concludes the paper with a brief discussion.

2. Correspondence and Gaussianization

2.1. Gaussian mixture model

In this paper, we break an image down into orderless N sub-image patches denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ where x_k is the kth patch of the image. Kinds of detectors can be used here to obtain sub-image patches and we adopt evenly sampled grid in our experiments. The basic idea of corresponding different scene images is to soft cluster all image patches across different images in an unsupervised manner, which is implemented through the use of a global Gaussian Mixture Model (GMM). This global GMM is trained on all patches across different images which provides a succinct description of the patch descriptor space by means of the M unimodal Gaussian components as well as their weights in the GMM.

The GMM is one of the most widely used statistical models for large-scale probability density estimation as it is capable of approximating any complex distributions at arbitrary precision with a sufficient number of Gaussian components.

Suppose the distribution of the image patches is modeled by a GMM of the form

$$p_X(x) = \sum_{k=1}^{M} w_k \mathcal{N}(x; \mu_k, \Sigma_k), \tag{1}$$

where w_k, μ_k and Σ_k denote the mixture weight, mean vector and covariance matrix of the kth Gaussian component, respectively, and M denotes the total number of Gaussian components.

This mixture density is a weighted linear combination of M unimodal Gaussian densities, namely,

$$\mathcal{N}(x; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}. \tag{2}$$

The unknown parameters of a GMM are collectively denoted as $\{w_k, \mu_k, \Sigma_k\}_{k=1}^{M}$. The maximum likelihood estimation of the GMM parameters is usually performed by using the Expectation–Maximization (EM) algorithm. In order to reduce the number of parameters to be estimated and to avoid overfitting, and in order to reduce the computational load for parameter estimation, the covariance matrices $\{\Sigma_k\}_{k=1}^{M}$ are restricted to be strictly diagonal (Reynolds et al., 2000). The use of diagonal covariance matrices for GMM learning has proven to be effective and computationally economical.

The EM algorithm for GMM training is an iterative process which proceeds as follows. At each iteration, the new parameter estimates guarantee that the data likelihood increases.

1. Start with an initialized parameter set:

$$\theta^{(0)} = \{w_k^{(0)}, \mu_k^{(0)}, \Sigma_k^{(0)}\}_{k=1}^{M}. \tag{3}$$

2. Given the training data set $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, at the nth iteration, compute the probability that the training vector x_t belongs to the Gaussian component k through Bayes rule:

$$p(k|x_t) = \frac{w_k^{(n)} \mathcal{N}(x_t; \mu_k^{(n)}, \Sigma_k^{(n)})}{\sum_{j=1}^{M} w_j^{(n)} \mathcal{N}(x_t; \mu_j^{(n)}, \Sigma_j^{(n)})}, \tag{4}$$

where $t = 1, 2, \ldots, N$, and $k = 1, 2, \ldots, M$.

3. Compute a new set of parameters:

$$w_k^{(n+1)} = \frac{1}{N}\sum_{t=1}^{N} p(k|x_t), \tag{5}$$

$$\mu_k^{(n+1)} = \frac{1}{N}\sum_{t=1}^{N} p(k|x_t)x_t, \tag{6}$$

$$\Sigma_k^{(n+1)} = diag\left\{\frac{1}{N}\sum_{t=1}^{N} p(k|x_t)x_t x_t^T\right\}. \tag{7}$$

4. If the parameter estimates converge, then stop. Otherwise, go to Step 2.

2.2. Multinomial trial

A GMM trained with the image patches from all the training images is obtained according to the previous section. Now, we present the process of extracting the proposed representation for each natural scene image.

We calculate the posterior probabilities of every patch of an image against all the Gaussian components in the GMM and use these posterior probabilities as the parameters for a multinomial trial, which randomly distributes the image patches into one of the Gaussian components.

The posterior probability that an observed patch x comes from the kth Gaussian component is given by

$$\gamma_k(x) = \frac{w_k\mathcal{N}(x;\mu_k,\Sigma_k)}{\sum_{j=1}^{M} w_j\mathcal{N}(x;\mu_j,\Sigma_j)}. \tag{8}$$

Now, we randomly distribute the observed patches into the M classes according to their posterior probabilities. That is, for a specific patch x, let $\gamma = [\gamma_1(x), \gamma_2(x), \ldots, \gamma_M(x)]^T$ be the parameters of a multinomial distribution $Mult(\gamma)$.

We introduce a M-dimensional binary random variable $\eta \sim Mult(\gamma)$. η is a 1-of-M vector representation in which a particular element η^k is equal to 1 and all other elements are equal to 0. $\eta^k = 1$ indicates that x is assigned to the kth class. Obviously, $p(\eta^k = 1) = w_k$ is the prior probability that x comes from the kth Gaussian component without knowing the actual value of x and $p(\eta^k = 1|x) = \gamma_k(x)$ is the posterior probability that x comes from the kth Gaussian component given the value of x.

Let random variable Y_k denote the samples in the kth class. Then Y_k has a probability density function given by,

$$p_{Y_k}(a) = p_{X|\eta}(a|\eta^k = 1), \tag{9}$$

$$= \frac{p(\eta^k = 1|a)p_X(a)}{p(\eta^k = 1)}, \tag{10}$$

$$= \frac{\gamma_k(a)p_X(a)}{w_k}, \tag{11}$$

$$= \mathcal{N}(a;\mu_k,\Sigma_k). \tag{12}$$

Now assume that a scene image has N patches. We separate them into M classes according to the above strategy. Let the number of samples in each class be denoted by n_k, where $k = 1, 2, \ldots, M$. We formulate a new random variable Z_k from Y_k as follows:

$$Z_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(Y_{ki} - \mu_k), \tag{13}$$

where Y_{ki} is the ith sample in the kth class. Since Y_{ki}'s are i.i.d, it is obvious that $Z_k \sim \mathcal{N}\left(0, \frac{\Sigma_k}{n_k}\right)$.

2.3. Gaussianized representation

A super-vector is formed according to the result of the multinomial trial described above. By normalizing and stacking the random variables Z_k's, we have a random vector

$$\widehat{Z} = \begin{pmatrix} \left(\frac{\Sigma_1}{n_1}\right)^{-\frac{1}{2}} Z_1 \\ \left(\frac{\Sigma_2}{n_2}\right)^{-\frac{1}{2}} Z_2 \\ \vdots \\ \left(\frac{\Sigma_M}{n_M}\right)^{-\frac{1}{2}} Z_M \end{pmatrix}. \tag{14}$$

It is straightforward to show that $\widehat{Z} \sim \mathcal{N}(0, I)$. $\widehat{Z}$ is a compact representation of a scene image which observes the standard normal distribution.

2.4. Correspondence

The orderless patch descriptors are not well corresponded across different images. In particular, natural scene images do not have rigid spatial correspondence as aligned face images. The multinomial trial establishes such correspondence by statistically assign each patch into different classes. The members of corresponding classes across images are corresponded. Therefore, the Gaussianized representation, which could be viewed as concatenated statistics from the classes, serves as a way to achieve correspondence for the orderless patches from different natural scene images.

3. Connection to histogram representation

Histogram representation, as a description for orderless patch-based features, has been widely used in visual recognition and image retrieval Schiele and Crowley (2000), Swain and Ballard (1991). For scene classification, histogram give the roughly alignment on patches by assigning each patch to one of the histogram bin. Moreover, such a representation can easily generate a similarity measure between two images based on the difference of the corresponding histograms.

Several approaches have been proposed in the literature to overcome the well-known limitations of the histogram representation, in particular, its sensitivity to outliers, choice of bins, and noise in the data. Soft assignment, which allows each feature vector belonging to multiple histogram bins, have been suggested to capture partial similarity between images Perronnin et al. (2006), Yang et al. (2008), van Gemert et al. (2008), Agarwal and Triggs (2006), Tuytelaars and Schmid (2007), Philbin et al. (2008). To enhance the discriminating capability of histograms, Farquhar et al. (2005) and Perronnin et al. (2006) introduced several ways to construct category-specific histograms, Larlus and Jurie (2006) and Yang et al. (2008) suggested to integrate histogram construction with classifier training, and Moosmann et al. (2007) proposed to use randomized forests to build discriminative histograms. Encoding high-dimensional feature vectors by a relatively small codebook, however, is the fundamental drawback that results in large quantization errors and lose of discriminability Boiman et al. (2008).

In the classical histogram-of-features representation, the histogram bins are chosen by a k-means algorithm on the whole patch data. Then each patch is distributed to a particular bin based on its distance to the cluster centroids. This process can be connected with the proposed Gaussianization in the following perspectives. First, k-means clustering leverages on the Euclidean distance, while the multinomial trial uses posterior on Gaussian mixtures, which leverages on the Mahamalobis distance. Second, k-means clustering provides deterministic bin membership while the Guassianized representation engages randomness via the multinomial trial parameters. Third, histogram-of-features only uses the number of patches assigned to histogram bins, while the proposed

representation also adopts the mean of features in each bin, leading to a more informative representation of the scene image.

Therefore, the proposed Gaussianized representation can be viewed as a generalized framework for the histogram-of-features. In particular, the multinomial trial adopts the Mahamalobis distance and alleviate the hard assignment issue of feature binning. Statistics are accumulated from patches randomly assigned to each bin or class, to take advantage of not only patch membership among the classes, but also a succinct description of patches within each class. We believe these characteristics contribute to the improved performance of natural scene categorization using the proposed Gaussianized representation, compared with the widely used histogram-of-features.

4. Patch-based representation

4.1. Patch extraction

In this section, we briefly describe the feature extraction process. According to the study by Fei-Fei and Perona (2005), the dense regular grid performs better than other sophisticated detectors for scene recognition. In this paper, we therefore choose to use the dense regular grid as our detector. More precisely, we use an evenly sampled grid spaced at 4×4 pixels, on which patches of 30×30 pixels are extracted from the scene images.

4.2. Patch descriptor

Two different representations for describing a patch are adopted, namely the raw features and the SIFT descriptors, respectively. The raw features and the SIFT descriptors are further transformed into Gaussianized super-vector representations by a global GMM trained from the raw features or the SIFT descriptors, respectively, as described Section 2. The following paragraphs give the details of the raw features and the SIFT descriptors.

The raw features consist of pixel intensities. We first resize a 30-by-30 patch to 6-by-6 and remove from the 6-by-6 patch the mean of the pixel intensity values, then normalize the intensity values to have a unit variance, and finally use the 2D discrete cosine transform (DCT) to generate the feature vector.

The SIFT descriptor is a 128-dimensional vector extracted from a 30-by-30 patch, whose dimensionality is reduced to 64 dimensions by PCA. SIFT stands for Scale-Invariant Feature Transform (SIFT) Lowe (1999) and is a widely used algorithm to detect and describe salient local features of an image. The SIFT features are local and based on the appearance at particular interest points, and are invariant to certain image transformations such as scaling and rotation. They are also robust to changes in illumination, noise, minor changes in viewpoint, as well as occlusion. The extraction of SIFT features consists of four major steps: (1) scale-space extrema detection, (2) keypoint localization, (3) orientation assignment, and (4) keypoint descriptor. In this paper, we compute the SIFT descriptor directly from the image patches. In particular, only the fourth step in the SIFT feature extraction process is necessary. That is, we compute a SIFT descriptor for each patch, based upon the histogram of gradients.

5. Experiments

5.1. Experiment setting

In this section, we investigate the effectiveness of our representation and further compare our results with existing works. We report our scene categorization experiment results on the scene category database which is composed of 15 scene categories, 13 provided by Fei-Fei and Perona (2005) and the other two collected by Lazebnik et al. (2006). Each scene category contains 200–400 images. The average size of the images is around 300×250 pixels. The major sources of the images include the COREL collection, personal photographs, and Google image search, etc. This database is one of the most comprehensive scene category databases used in the literature. Example images of different scene categories of this database are illustrated in Fig. 1.

Here, the experiment setting is made the same as that in (Fei-Fei and Perona, 2005 and Lazebnik et al., 2006) to guarantee the fairness of performance comparison. Specifically, all experiments are repeated ten times with 100 randomly selected images per class for training and the rest for testing and the average of per-class recognition rate is recorded for each run. The final result is reported as the mean and standard deviation of the results from the individual runs. We perform all processing in grayscale, even when color information is available.

We adopt two kinds of patch descriptors as described in Section 4. First, the DCT-based feature or the 128-dimensional SIFT vector is extracted within a 20×20 patch over a grid with spacing of five pixels. Then the dimension of the descriptor is reduced to 64 by Principal Component Analysis (PCA). The GMM used here containing 512 Gaussian components. When doing the calculation, instead of explicitly carrying out the multinomial trial, we can approximate according to Eq. (14) as $Z_k = \sum_{i=1}^{N} \gamma_{ik}(x_i - \mu_k)/\sum_{i=1} N n \gamma_{ik}$.

To demonstrate the effectiveness of the proposed representation, we employ three different classifiers for multi-class classification, namely, nearest neighbor (NN), nearest centroid (NC) and support vector machine (SVM). For the nearest neighbor classifier, we calculate the Euclidean distances of each image in the test set to all the images in the training set and assign to the test image the label of the nearest training image. For the nearest centroid classifier, we estimate the centroid of each class on the training set, and calculate the Euclidean distances of each test image to all the centroids and assign to the test image the label of the nearest centroid. For SVM, we use the LIBSVM software Chang and Lin, 2001 to perform training and testing. The kernel used is

$$K(\widehat{Z}_a, \widehat{Z}_b) = \frac{\widehat{Z}_a^T \widehat{Z}_b}{\sqrt{\|\widehat{Z}_a\| \|\widehat{Z}_b\|}}. \tag{15}$$

Note that in Eq. (15), we can first normalize each super-vector $\widehat{Z}$ as $M = \frac{\widehat{Z}}{\sqrt{\|\widehat{Z}\|}}$, then the kernel can be simplified as,

$$K(M_a, M_b) = M_a^T M_b \tag{16}$$

which is the dot-product of two vectors corresponding to two images. Here, we use the linear kernel as it is efficient in the evaluation stage and easy to be deploy in large-scale applications.

5.2. Experiment results

5.2.1. Thirteen-category scene classification

Table 1 shows the results of our representation on 13 categories scene classification experiments, which has the same 13 categories and experiments setting as used in (Fei-Fei and Perona, 2005) for fair comparison. Here we present the performance under the different numbers of Gaussian components and different features. The GMM used for computing the super-vector $\widehat{Z}$ for each scene image contains 512 or 1024 Gaussian components. The raw features and SIFT features are adopted, respectively. Among the three classifiers, the SVM classifier achieves better performance than the other two classifiers. Note that even with the raw features, the SVM classifier of our representation achieves an average classification accuracy of 74.4%, which is much higher than the best

Fig. 1. Example images from the scene category database.

Table 1
Classification results on the 13 scene category database.

Mixture number	Raw feature			SIFT		
	KNN	NC	SVM	KNN	NC	SVM
512	60.3 ± 0.8	66.5 ± 1.0	**72.6** ± 0.9	73.6 ± 0.8	78.3 ± 0.7	**83.6** ± 0.6
1024	61.5 ± 0.7	69.0 ± 0.9	**74.4** ± 1.0	74.0 ± 0.6	78.7 ± 0.6	**84.1** ± 0.5

recognition rate (65.2%) in (Fei-Fei and Perona, 2005), obtained with SIFT and the histogram representations method. With the SIFT descriptors and SVM classifier, our system achieves the average classification accuracy of 84.1%.

5.2.2. Fifteen categories scene classification

We also perform our representation on 15 categories scene classification experiments. Note that the features used here is slightly different from that in the previous experiments. Besides the 64 dimension SIFT descriptor obtained by PCA, the coordinate information (x, y coordinates) for each patch is used as two additional dimensions to form 66 dimensional features. The coordinates used here can help to capture the spatial information of the patches, and further improve scene categorization performance.

Table 2 compared the average classification accuracy by our proposed representation with the traditional histogram representation. It is clear to see that our proposed representation greatly outperform the histogram representation regardless of the underly classifiers chosen. In most of the categories, SVM shows better performance than the other two classifiers. On average, both representations achieved the best performance when adopting SVM as the classifier, and the Gaussianzation representation is 20% better than histogram representation in that situation.

We can further compare our results with existing work that goes beyond orderless bag-or-feature representation. In (Lazebnik et al. (2006)), Lazebnik et al. introduced spatial pyramid matching (SPM) to incorporate the spatial information with histogram representation and reported an accuracy of 81.4% on 15 categories. In the experiment, our new representation achieves a superior performance of 83.5% in accuracy without using the SPM strategies.

Fig. 2 shows the confusion patterns between the 15 scene categories by the histogram representation and the proposed representation, respectively. By comparing the diagonal elements, we can find that our new representation outperforms histogram representation on all 15 categories. The highest recognition rate is obtained for the reign of "Calsurburb" and "PARoffice" for our new representation. The highest misidentified rate is that 19% of "livingroom" is

Table 2
Classification results on the 15 scene category database.

Category	Histogram			Gaussianization		
	NC	NN	SVM	NC	NN	SVM
CALsuburb	43.3 ± 2.4	75.7 ± 4.2	89.1 ± 2.0	99.2 ± 0.9	92.1 ± 2.0	99.7 ± 0.4
MITcoast	32.9 ± 2.6	38.5 ± 3.3	73.2 ± 3.1	83.1 ± 3.1	75.4 ± 3.0	87.9 ± 2.2
MITforest	92.1 ± 0.8	87.9 ± 2.3	93.9 ± 0.9	96.5 ± 1.0	95.5 ± 1.6	96.1 ± 1.6
MIThighway	70.4 ± 2.7	69.6 ± 2.1	79.6 ± 1.8	82.0 ± 2.2	85.8 ± 2.4	91.1 ± 1.5
MITinsidecity	47.6 ± 3.6	31.2 ± 3.0	62.7 ± 2.9	79.3 ± 1.9	67.6 ± 2.9	85.5 ± 1.6
MITmountain	29.3 ± 2.0	47.7 ± 3.1	79.6 ± 3.6	84.6 ± 2.1	72.0 ± 1.8	91.9 ± 1.9
MITopencountry	36.0 ± 2.3	44.8 ± 3.2	59.1 ± 2.6	74.3 ± 2.1	57.7 ± 3.6	75.6 ± 1.8
MITstreet	42.1 ± 1.7	960.6 ± 2.0	75.1 ± 2.1	84.9 ± 2.4	75.2 ± 1.7	89.4 ± 1.8
MITtallbuilding	31.6 ± 3.4	35.9 ± 2.4	80.4 ± 2.7	84.2 ± 1.2	75.0 ± 2.4	91.4 ± 1.5
PARoffice	42.1 ± 2.5	55.1 ± 4.6	78.4 ± 2.1	95.8 ± 1.4	89.0 ± 2.9	96.3 ± 1.7
Bedroom	19.0 ± 3.7	17.4 ± 3.6	32.8 ± 3.0	61.1 ± 5.5	41.4 ± 4.6	68.5 ± 5.2
Industrial	10.1 ± 1.1	22.4 ± 2.6	24.9 ± 2.5	46.6 ± 1.8	45.9 ± 3.7	65.0 ± 2.8
Kitchen	50.0 ± 5.6	34.5 ± 3.7	50.6 ± 5.8	68.5 ± 4.8	49.5 ± 4.5	76.1 ± 4.1
Livingroom	27.2 ± 3.5	30.2 ± 3.8	30.6 ± 4.1	54.1 ± 4.5	28.5 ± 3.3	58.5 ± 3.0
Store	77.0 ± 2.3	41.7 ± 5.3	63.3 ± 4.4	70.5 ± 1.8	51.5 ± 2.7	78.8 ± 2.3
Average	43.4 ± 1.0	46.2 ± 0.9	64.9 ± 0.5	77.7 ± 0.6	66.8 ± 0.5	83.5 ± 0.3

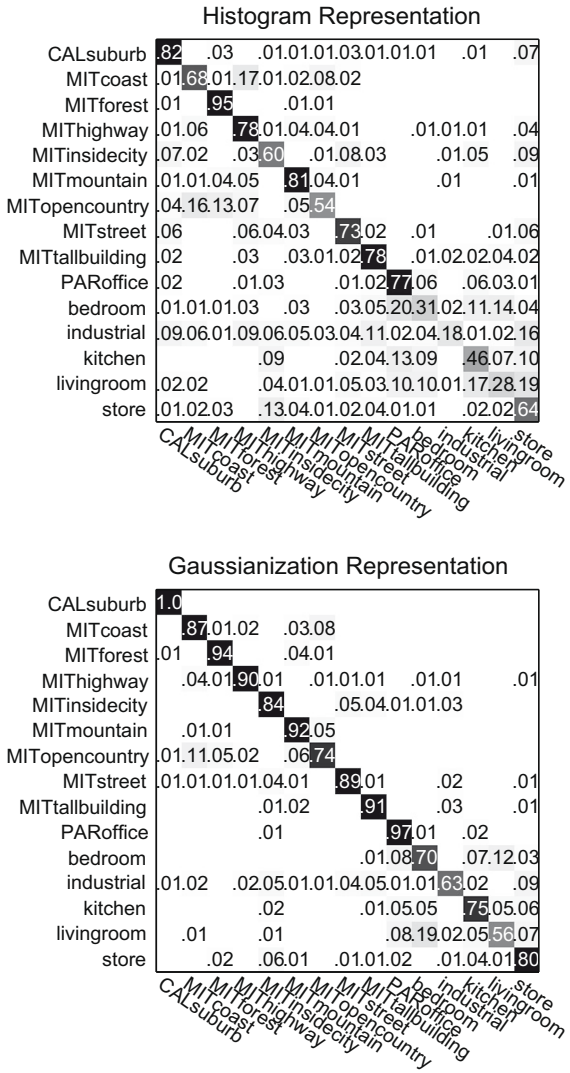

Fig. 2. Comparison confusion matrices on scene category database for histogram representation and Gaussianization representation. The entry in the *i*th row and *j*th column is the percentage of images from class *i* that were misidentified as class *j*. For better viewing, please see the pdf file.

misidentified as "bedroom", which is reasonable given the highly similar configuration in the two categories.

6. Conclusion and discussion

In this paper, we propose a Gaussianization vector representation for natural scene categorization, which represents a scene image as a super-vector observing the standard normal distribution. We apply various classification techniques on this representation of feature vectors and achieve significantly improved performance on scene categorization as compared with previous work using the popular histogram-of-features representation. In particular, our experiments show that this representation, without considering the spacial information, achieves much better performance than the bag-of-words with pLSA (Fei-Fei and Perona, 2005). Furthermore, it outperforms the bags of features with spatial pyramid matching approach in (Lazebnik et al., 2006).

Acknowledgement

This work was supported in part by the US Government VACE program and in part by the National Science Foundation Grants CCF 04-26627, 08-03219 and 08-07329. The results and conclusions expressed in this paper are those of the authors, and are not endorsed by the NSF.

References

Agarwal, A., Triggs, B., 2006. Hyperfeatures-multilevel local coding for visual recognition. In: Lecture Notes in Computer Science.
Boiman, O., Shechtman, E., Irani, M., 2008. In defense of nearest-neighbor based image classification. In: CVPR.
Chang, C.-C., Lin, C.-J., 2001. LIBSVM : A library for support vector machines. In: Software available at <http://www.csie.ntu.edu.tw/cjlin/libsvm>.
Farquhar, J., Szedmak, S., Meng, H., Shawe-Taylor, J., 2005. Improving bag-of-keypoints image categorisation. In: Technical Report.
Fei-Fei, L., Perona, P., 2005. A Bayesian hierarchical model for learning natural scene categories. In: CVPR.
Grauman, K., Darrell, T., 2005. Pyramid match kernels: Discriminative classification with sets of image features. In: ICCV.
Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 41, 177–196.
Jiao, F., Li, S., Shum, H., Schuurmans, D., 2003. Face alignment using statistical models and wavelet features. In: CVPR.
Larlus, D., Jurie, F., 2006. Latent mixture vocabularies for object categorization. In: British Machine Vision Conference.
Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR.
Lowe, D., 1999. Object recognition from local scale-invariant features. In: Proc. IEEE Internat. Conf. on Computer Vision, pp. 1150–1157.

Moosmann, F., Triggs, B., Jurie, F., 2007. Randomized clustering forests for building fast and discriminative visual vocabularies. In: NIPS.
Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. Internat. J. Comput. Vision 42.
Perronnin, F., Csurka, G., Dance, C., Bressian, M., 2006. Adapted vocabularies for generic visual categorization. In: European Conference on Computer Vision.
Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR.
Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted gaussian mixture models. Digit. Signal Process. 10 (1-3), 19–41.
Schiele, B., Crowley, J., 2000. Recognition without correspondence using multidimensional receptive field histograms. IJCV 36 (1), 31–50.
Swain, M., Ballard, D., 1991. Color indexing. IJCV 7 (1), 11–32.
Treisman, A., Gelade, G., 1980. A feature-integration theory of attention. Cognitive Psychol. 12, 97–136.
Tuytelaars, T., Schmid, C., 2007. Vector quantizing feature space with a regular lattice. In: ICCV.
van Gemert, J.C., Geusebroek, J., Veenman, C., Smeulders, A., 2008. Kernel codebooks for scene categorization. In: European Conf. on Computer Vision.
Vogel, J., Schiele, B., 2004. A semantic typicality measure for natural scene categorization. In: DAGM'04 Annual Pattern Recognition Symposium.
Wallraven, C., Caputo, B., Graf, A., 2003. Recognition with local features: The kernel recipe. In: ICCV, vol. 1, pp. 257–264.
Wang, W., Shan, S., Gao, W., Cao, B., Yin, B., 2002. An improved active shape model for face alignment. Multimodal Interfaces.
Willamowski, J., Arregui, D., Csurka, G., Dance, C.R., Fan, L., 2004. Recognition with local features: The kernel recipe. In: ICPR Workshop on Learning for Adaptable Visual Systems.
Yang, L., Sukthankar, Rahul, Jin, R., Jurie, F., 2008. Unifying discriminative visual codebook generation with classifier training for object category recognition. In: CVPR.
Zhou, X., Zhuang, X., Tang, H., Hasegawa-Johnson, M., Huang, T.S., 2008. A novel Gaussianized vector representation for natural scene categorization. In: ICPR.