A supervised LPP and Neural Network based Scene classification with Color Histogram and Camera Metadata

Xian-Hua HAN[†] Yen-Wei CHEN[†] Hideto FUJITA[‡] Kan-Ichi KOYAMA[‡] and Wataru TAKAYANAKI[‡]

[†] College of Information Science and Engineering, Ritsumeikan University
[‡] Digital Technology Research Center, SANYO Electric Co., Ltd.
E-mail: [†] hanxhua@fc.ritsumei.ac.jp

Abstract Scene classification (e.g., landscape, sunset, night-landscape, etc.) is still a challenging problem in computer vision. Scene classification based only on low-level vision cues has had limited success on unconstrained image sets. In other hand, camera metadata related to capture conditions provides cues independent of the captured scene content that can be used to improve classification performance. Analysis of camera metadata statistics for images of each class revealed that some metadata fields are most discriminative for some classes. So, in this paper, we proposed to use the combined feature of scene color histogram and camera metadata, and then using supervised Locality preserving projection(LPP) for feature space transformation and dimension reduction, and finally, adapt Probabilistic neural network for scene classification. Experimental results show that the classification accuracy rate can be improved compared with using PCA (Principal Component Analysis) subspace learning method, and are also better than that with only the low-level vision feature(color histogram).

Keyword Windows, Word, Technical Report, Template

1. Introduction

Automatically determining the semantic classification (e.g., sunset, flower, landscape) of an arbitrary image is a difficult problem [1,2]. Much research has been done recently, and a variety of classifiers and feature sets have been proposed. The most common design for such systems has been to use low-level features (e.g., color, texture) and statistical pattern recognition techniques [3,4], and achieve some success in constrained (e.g., Corel) environments. Meanwhile, the recent proliferation of digital images has created a greater need for scene classification and at the same time, provided an additional source of information to help solve the problem: camera metadata embedded in the image files. Metadata recorded by the camera includes information related to the image capture conditions and values such as presence or absence of flash, subject distance and exposure value. So in this paper, we proposed to use the combined feature of the low-level feature (color histogram) and camera metadata for scene classification. However, The color space is usually of high dimension. the high dimensionality of feature vectors results in high computational cost. Several transformations, including principal component analysis (PCA) have been proposed to reduce the dimensionality [5]. However, PCA is an efficient method for represent information of data, but is usually non-effective for classification, especially for complicated data like the images in scene classification problem.

Recently, Locality Preserving Projections (LPP) was proposed for approximate the eigenfunctions of the Laplace Beltrami operator on the Face manifold[6] and new test can be explicitly mapped to the learned low-dimensional sub-manifold. Different from subspace learning representation PCA which is optimal in the sense of global Euclidean structure, LPP is optimal in the sense of local manifold structure. However, LPP uses the nearest-neighbor graph to seek the nearest neighbors, which will fail on database containing complex variations, because the nearest neighbor samples may belong to different classes in databases containing complex variations. So in this paper, we proposed to use supervised LPP for learning the scene subspace which can not only find true manifold of images but also recover the intrinsic geometric structure of image feature space for scene classification. After obtain subspace manifold, we use Probabilistic Neural Network to handle the subspace feature for scene classification. The Probabilistic Neural Network (PNN) is a kind of neural networks that tries to resolve classification problems (i.e. the separation of the input space into different regions) that relies on the Bayes decision theory rule applied to a multidimensional input space. This architecture uses one unit for each training pattern and a one-pass learning process, that is independent from other patterns. As a consequence, it is very fast to train and easily extendable with new training patterns. Application can be found, for example, in the area of real-time classification [3]. Experimental results show that the accuracy rate of classification with our proposed method can be greatly improved compared with other methods.

2. The image feature

2.1 Color histogram

Color histograms are widely used to capture the color information in an image. They are easy to compute and tend to be robust against small changes of camera viewpoints. Given an image **I** in some color space (e.g., red, green, blue). The color channels are quantized into a coarser space with *k* bins for red, *m* bins for green and *l* bins for blue. Therefore the color histogram is a vector $\mathbf{h}=[h_1,h_2,\cdots\cdots,h_n]^T$, where *n*=k*m*l, and each element h_i represents the number of pixels of the discretized color in the image. We assume that all images have been scaled to the same size. Otherwise, we normalize histogram elements as:

$$h_j^{'}=\frac{h_j}{\sum_{j=1}^{j=n}h_j} \qquad (1)$$

$\mathbf{h}=[h_1^{'},h_2^{'},\cdots\cdots,h_n^{'}]^T$ is the normalized color histogram and is as the feature vector to be stored as the index of the image database.

2.2 Digital camera metadata

The specification for camera metadata (used for JPEG images) includes hundreds of tags. Among these, 26 relate to picture taking conditions (e.g., Subject Distance, scene Brightness(bv-value), FlashUsed,and ExposureTime). It is clear that some of these cues can help distinguish various classes of scenes. For example, low scene Brightness tends to be appear more frequently with night images(example for night-landsacep, candle light) than other types of images. Some tags will be more useful than others for a given problem.

In this paper, we use two of these tags for assisting classification that we believe to be useful for scene classification of our database. One is the bv-value which represents global brightness of the camera images. With the statistical analysis of the bv-value, it is very clear that most of candle-light and night-landscape scenes have minus value, whereas the other types scenes including landscape, flower, sunset, text and back-light. So in this paper, the bv-value can be used to separate our scene database into two groups: candle-light, night-landscape scenes group and the other types scenes group. At the same time, we also can consider the bv-value as the one feature of all scene features, and then use our subspace learning and PNN method for classification. The other camera tag is the subject distance which represent the distance from the camera to the subject. With few exceptions, only sunset, night-landscape and landscape scenes in particular, can have a large subject distance and a part of backlight scenes also have large subject distance. Therefore, we expect distance measures to discriminate strongly between the above mentioned types scenes and other scenes including text and flower. This subject distance is consider as one feature of all scene features.

3. Supervised Locality Preserving projections

3.1 The Problem

The problem of subspace learning for image indexing and representation is the following. Given a set of scene feature space $\mathbf{x}_1,\mathbf{x}_2,\cdots\cdots,\mathbf{x}_m$ in $\mathbf{R}^n$ of images, the goal is to find a lower dimensional feature representation $\mathbf{f}_i$ of $\mathbf{x}_i$ such that $||\mathbf{f}_i-\mathbf{f}_j||$ reflects the neighborhood relationship between $\mathbf{f}_i$ and $\mathbf{f}_j$. In other word, if $||\mathbf{f}_i-\mathbf{f}_j||$ is small, then $\mathbf{x}_i$ and $\mathbf{x}_j$ are belong to same class. Here, we assume that the images reside on a sub-manifold embedded in the ambient space $\mathbf{R}^n$.

3.2 The Algorithm

In this section, we give a brief description of Locality Preserving Projections (LPP)[3]. LPP seeks a linear transformation P to project high-dimensional data into a low-dimensional sub-manifold thatpreserves the local Structure of the data. $\mathbf{x}_1,\mathbf{x}_2,\cdots\cdots,\mathbf{x}_m$ denote the set of scene feature of image sample vectors in $\mathbf{R}^n$ and $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots\cdots,\mathbf{x}_m]$ denotes the color histogram matrix whose column vectors are histograms. The linear transformation **P** can be obtained by solving the following minimization problem:

$$\min_{\mathbf{P}} \sum_{ij} (\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j)^2 B_{ij} \qquad (2)$$

where B_{ij} evaluates the local structure of the image space.

In this paper, we use normalized correlation coefficient of two sample as the penalty weight if the two sample belong to the same class:

$$B_{ij} = \left\{ \begin{array}{ll} \dfrac{\mathbf{x}_i{}^T\mathbf{x}_j}{\sum_l^n x_{il}{}^2 \sum_l^n x_{jl}{}^2}, & if\ sample\ i\ and\ j\ is\ in\ same\ class \\ 0, & otherwise \end{array} \right\} (3)$$

By simple algebra formulation, the objective function can be reduced to:

$$\begin{aligned} &\frac{1}{2}\sum_{ij}(\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j)^2 B_{ij} \\ &= \sum_i \mathbf{P}^T\mathbf{x}_i D_{ii} \mathbf{P}^T\mathbf{x}_i - \sum_{ij} \mathbf{P}^T\mathbf{x}_i B_{ij} \mathbf{P}^T\mathbf{x}_j \\ &= \mathbf{P}^T\mathbf{X}(\mathbf{D}-\mathbf{B})\mathbf{X}^T\mathbf{P} = \mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P} \qquad (4) \end{aligned}$$

D is a diagonal matrix; its entries are column (or row, since **B** is symmetric) sums of **B**, $D_{ii} = \sum_j B_{ij}$. **L**=**D**-**B** is the Laplacian matrix [5]. Then, The linear transformation P can be obtained by minimizing the objective function under constraint:

$$\mathbf{P} = \arg\min_{\mathbf{P}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P}=1} \mathbf{P}^T\mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T\mathbf{P} \qquad (5)$$

Finally, the minimization problem can be converted to solving a generalized eigenvalue problem as follows:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P} = \lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P} \qquad (6)$$

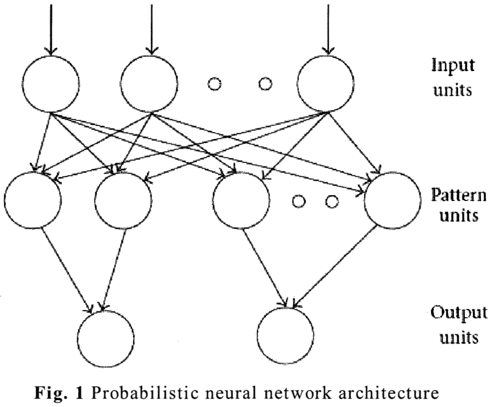


Fig. 1 Probabilistic neural network architecture

4. Probabilistic neural network

The PNN model is based on Parzen's results on probability density function (PDF) estimators [7, 8]. PNN is a three-layer feedforward network consisting of input layer, a pattern layer, and a summation or output layer as shown in Figure 1. We wish to form a Parzen estimate based on *K* patterns each of which is *n*-dimensional, randomly sampled from *c* classes. The PNN for this case consists of *n* input units comprising the input layer, where each unit is connected to each of *K* pattern units; each pattern unit is, in turn, connected to one and only one of the *c* category units. The connection from the input to pattern units represents modifiable weights, which will be trained. Each category unit computes the sum of the pattern units connected to it. A radial basis function and a Gaussian activation function are used for the pattern nodes.

The PNN is trained in the following way. First, each pattern(sample feature) **f** of the training set is normalized to have unit length. The first normalized training pattern is placed on the input units. The modifiable weights linking the input units and the first pattern unit are set such that $\mathbf{w}_1 = \mathbf{f}_1$. Then, a single connection from the first pattern unit is made to the category unit corresponding to the known class of that pattern. The process is repeated with each of the remaining training patterns, setting the weights to the successive pattern units such that $\mathbf{w}_k = \mathbf{f}_k$ for *k* = 1, 2, . . . ,*K*. After such training, we have a network which is fully connected between input and pattern units, and sparsely connected from pattern to category units. The trained network is then used for classification in the following way. A normalized test pattern **f** is placed at the input units. Each pattern unit computes the inner product to yield the net activation **y**.

$$\mathbf{y}_k = \mathbf{w}_k{}^T * \mathbf{f} \qquad (7)$$

and emits a nonlinear function of $\mathbf{y}_k$; each output unit sums the contributions from all pattern units connected to it. The activation function used is $\exp^{(\|\mathbf{x}-\mathbf{w}_k\|/\delta^2)}$. Assuming that both **x** and $\mathbf{w}_k$ are normalized to unit length, this is equivalent to using $\exp^{(\|\mathbf{x}-1\|/\delta^2)}$.

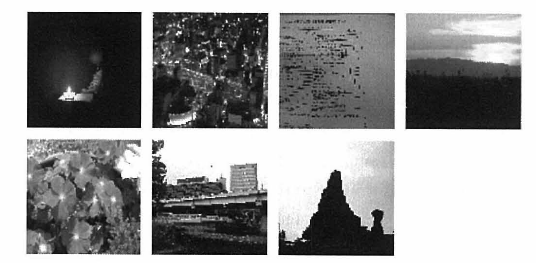
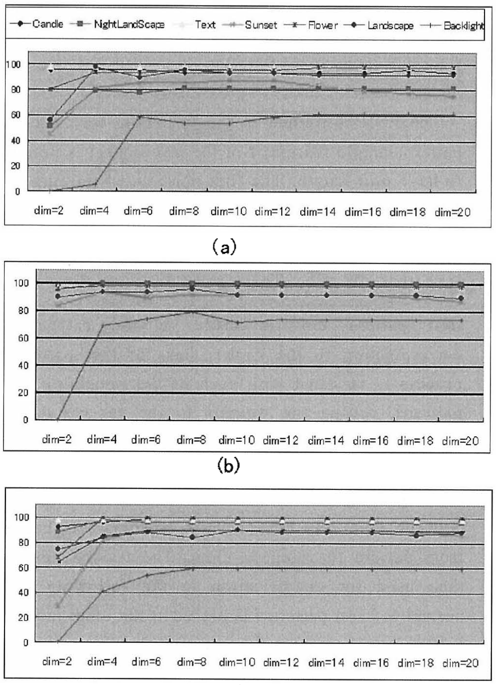


Fig. 2 sample scenes.



(a)

(b)

(c)

Fig. 3 The accuracy rate Vs. dimention by SLPP feature extraction and PNN classifier; (a)The accuracy rate with only color features(LPP(Color)); (b)The accuracy rate with Combined features(LPP(Arch1)); (b)The accuracy rate with Combined features(LPP(Arch2)).

In This paper, we used the combinated features of color histogram and camera metadata as the input elements of PNN. In the first architecture, the scene database is firstly separated into two groups according to the camera BV-value, and then use supervised LPP subspace learning method for feature extraction with color histogram and camera distance features, and finally apply the LPP subspace features to PNN for scene classification; The second architecture combine the two camera feature(BV-value and subject distance) and color histogram into scene global feature, and then use supervised LPP for subspace feature extraction and PNN for classification of all scene classes.

5. Experimental results

Scene classification experiments are carried out to compare the proposed method with other low-dimensional feature indices. Most of the training scenes and the test scenes have the size 2816 by 2112.The database includes seven classes: landscape, night-landscape, sunset, candle light, flower, back-light, text(Fig.2 give one scene for each class in sequence). each class has about 50 scenes(altogether 337 scenes). In our experiments, we use Leave-one-out method, each scene is as a test scene one time and the others are as training scene for subspace basis functions extraction and PNN training. We discretize each RGB color channel to 8 levels. Therefore, the color histogram feature vector has 512 components. In the first architecture(denoted by SLPP(Arch1)), the combined features(513 components) of subject distance and color histogram are used for subspace learning, and M(2,4,6,...,20) SLPP basis functions are retained for feature extraction. So the obtained M subspace features after projection are as PNN inputs for training. The output of the first group PNN has 2 units(represent candle light and night-landscape, respectively) and output of the first group PNN has 5 units(represent text, sunset, flower and so on).In the second architecture(denoted by SLPP(Arch2)), the combined features(514 components) of two camera tags and color histogram are used for subspace learning, and also M(2,4,6,...,20) SLPP basis functions are retained for feature extraction. So the obtained M subspace features after projection are as PNN inputs for training. The output of PNN has 7 units and each output unit represent each scene class, respectively. We also give experimental results using SLPP feature extraction and PNN learning only with 512 color features(color histogram) for comparison(denoted by SLPP(Color)). Figure 3 show the accuracy rate of different types of test scenes (Candle:50;

Night-LandScape:49; Text:50; Sunset:49; Flower:50; Landscape:50; Backlight:39; altogether: 337 test scenes) of different algorithms(fig.3(a) LPP(color); fig.3(b) SLPP(Arch1); fig.3(c) SLPP(Arch2)) vs. dimensions(Retained LPP features). It is clear from Fig. 3 that the accuracy of backlight scenes is about lower than 60%, and the accuracy rate of night-landscape and sunset scenes are also not high and just near to 80% in the SLPP(Color) algorithm (without metedata information); SLPP(Arch2) can improve the accuracy of night-landscape and sunset scenes classes to about 90%, but the accuracy of backlight scenes is similar to the one of SLPP(Color); SLPP(Arch1) can not only improve the accuracy rate of night-landscape and sunset scenes to more than 90%, but also greatly enhance the accuracy rate of backlight to about 80%. Figure 4 give the average accuracy rate of all(7) scene classes vs. dimensions(dimension=2,4,6,...,20).

In order to validate SLPP features more efficient than other subspace learning algorithm(example for PCA--Principal Component Analysis) in classification field, we also applied PCA for feature extraction(PNN learning for classification) for color histogram(PCA(Color), two groups of scene selection with 513 combined features of color histogram and metedata (PCA(Arch1)) and 514 combined features (PCA(Arch2)). The compared accuracy rates of different scene classes are show in Fig. 5(10 PCA features and 8 SLPP features). It can be seen from Fig. 5 that the accuracy rate of each scene class with SLPP feature extraction is higher than that using PCA feature extraction. Figure 6 gives the average accuracy rate of all 337 test scenes using PCA and SLPP feature extraction and PNN classification. It is evident that SLPP can extract more efficient feature than PCA for classification, and the accuracy rate can be greatly improved only combining a few camera metadatas (here only two metedatas) with low-level visual feature of scenes.

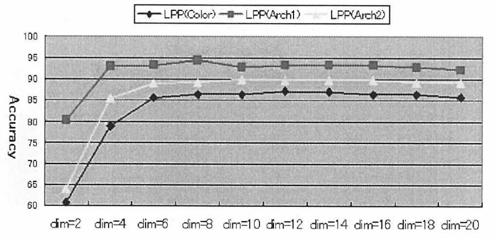


Fig. 4 The average accuracy rate of all test scenes using LPP(Color) LPP(Arch1) and LPP(Arch2).

6. Conclusions

In this paper, we proposed to combine camera metadata and color histogram as scene classification feature, and at the same time, applied supervised LPP to extract a new index from the original feature space, and then used PNN for classification. Experiment results showed that the accuracy rate of classification can be greatly improved compared with other low-dimensional feature indices (example for PCA).

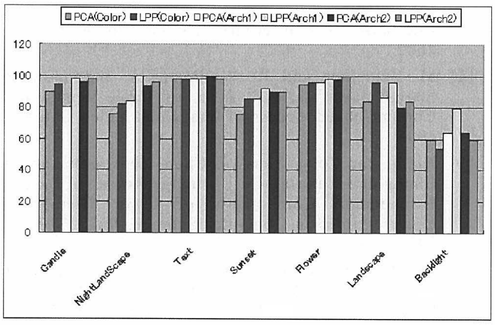


Fig. 5 The compared accuracy rate of different scene types using PCA and LPP for feature extraction, respectively.

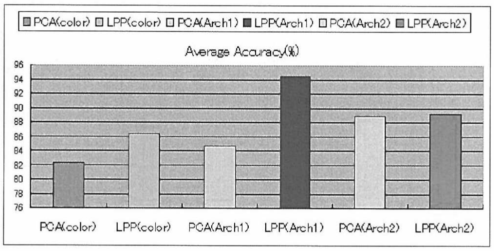


Fig. 6 The compared average accuracy rate of all test scenes using PCA and LPP for feature extraction, respectively.

Reference

[1] Lalit Gupta, Vinod Pathangay, Arpita Patra, A. Dyana, and Sukhendu Das, "Indoor versus Outdoor Scene Classification Using Probabilistic Neural Network," EURASIP Journal on Advances in Signal Processing Volume 2007 (2007), Article ID 94298, 10 pages.

[2] Andrew Payne and Sameer Singh, " Indoor vs. outdoor scene classification in digital photographs, " Pattern Recognition Volume 38, Issue 10, October 2005, Pages 1533-1545.

[3] N. Serrano, A. Savakis, and J. Luo, " A Computationally Efficient Approach to Indoor/Outdoor Scene Classification," Proc. of International Conference on Pattern Recognition, 2002.

[4] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," Proc. of IEEE Workshop on Content-based Access of Image and Video Databases, 1998.

[5] Xiang-Yan Zeng, Yen-Wei Chen, Zensho Nakao, Jian Cheng and Hanqing Lu, "Image Retrieval Based on Independent Components of Color Histograms," Lecture Notes in Computer Science, Vol. 2773/2003, 1435-1442.

[6] Xiaofei He and Partha Niyogi, "Locality Preserving Projections," Advances in Neural Information Processing Systems 16, Vancouver, Canada, 2003.

[7] D. F. Specht,"Probabilistic neural networks, " Neural Networks, vol. 3, no. 1, pp. 109.118, 1990.

[8] P. E. H. Richard, O. Duda, and D. G. Stork, "Pattern Classification," JohnWiley & Sons, New York, NY, USA, 2004. Hindawi