JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 33, 1177-1193 (2017)
DOI: 10.6688/JISE.2017.33.5.5

Image Classification Using Naive Bayes Classifier With Pairwise Local Observations*

SHIH-CHUNG HSU[1], I-CHIEH CHEN[1] AND CHUNG-LIN HUANG[2,+]
[1]*Department of Electrical Engineering*
National Tsing-Hua University
Hsinchu, 300 Taiwan
[2]*Department of M-Commerce and Multimedia Applications*
Asia University
Taichung, 413 Taiwan
E-mail: clhuang@asia.edu.tw

We propose a pairwise local observation-based Naive Bayes (*NBPLO*) classifier for image classification. First, we find the salient regions (*SRs*) and the Keypoints (*KPs*) as the local observations. Second, we describe the discriminative pairwise local observations using Bag-of-features (*BoF*) histogram. Third, we train the object class models by using random forest to develop the *NBPLO* classifier for image classification. The two major contributions in this paper are multiple pairwise local observations and regression object class model training for *NBPLO* classifier. In the experiments, we test our method using Scene-15 and Caltech-101 database and compare the results with the other methods.

Keywords: local observation-based Naive Bayes classifier (*NBPLO*), salient region (*SR*), keypoint (*KP*), bag-of-feature (*BoF*), image clssification

1. INTRODUCTION

Image classification has been a challenging unsolved problem due to the complexity of image contents. It has been a popular research subject of many recently published researches [1-6] using the image datasets such as Scene-15 [5], Pascal VOC [6], the Caltech-101 [2], Sun Database [3], Caltech-256 [8], and ImageNet [7]. ImageNet [7] is the most challenging dataset with ever increasing number of image categories which has become a publicly available large-scale benchmark data for image classification contest. The image classification accuracy is determined by effective feature extraction and classifier training.

Most of the image classification methods are based on the local features and the global features. They consist of mostly the edge-based feature points such as *SIFT* [9], *SURF* [10], Harris corner [11], *HOG* [12], and dense *SIFT* [13]. The global feature algorithms aim to recognize the image content as a whole, however, they often are not related to high-level semantics. The local feature algorithms focus mainly on keypoints and the salient image patches. The *SIFT* is a promising low-level visual descriptor which has been used as the basis of a bag-of-visual words (*BVW*) model [14, 15]. The *BVW* is a promising method for visual content classification, however, it usually describes visual

Received April 6, 2016; revised December 29, 2016; accepted February 17, 2017.
Communicated by Tyng-Luh Liu.
+ Corresponding author.
* This research was financially supported by MOST of Taiwan under the research grant No. MOST 103-2221-E468-006-MY2.

data at a non-semantic level. Based on the probability distribution [16] and soft weighting [17] of *BVW*, Kesorn *et al.* [18] propose an ontology-based model to bridge the low-level feature and high-level semantic concept for object/scene recognition. Sánchez *et al*. [19] use the Fisher Kernel framework as an alternative BVWs representation with generative Gaussian mixture model.

To categorize the images, the histogram-like bag-of-features (*BoF*) [20, 21] extracted from local patches can be proposed for image representation. The *BOF*s are encoded by using various algorithms such as K-means (or VQ) and sparse coding (SC). Then the local features are grouped to provide global semantics. The *BoF* methods disregard the information about the spatial layout of the features. Hence, they have severely limited global description capability. The spatial-pyramid-matching (*SPM*) method [22] is proposed to replace the *BoF* method which is an orderless collection of local features. The *SPM* method partitions the image into increasing fine sub-regions and computes the histograms of each sub-region as the local feature.

Lazebnik *et al.* [1] quantize the *SIFT* feature and the oriented edge points respectively for *SPM* to approximate the global geometric correspondence. Yang *et al*. [24] extend the *SPM* method by generalizing VQ to sparse coding and proposing a linear *SPM* kernel based on *SIFT* sparse codes. Wang *et al*. [4] present the Locality-constrained Linear Coding (*LLC*) in *SPM*. The *LLC*-based method utilizes the locality constraints to project each descriptor into its local-coordinate system and integrate these descriptors to generate the final representation. Jiang *et al*. [26] present a label consistent K-SVD (LC-KSVD) algorithm for discriminative sparse coding dictionary learning and introduce a new label consistency constraint to form a unified objective function. Uijlings *et al.* [27] raise the question regarding how to use the spatial contents to recognize the objects. They point out that using random sampling [28] and regular dense grid [29] of local patches show better performance than using the interest points [30]. Jia *et al.* [31] also question the effectiveness of using *SPM*-based feature for image classification and examine the effectiveness of receptive field designs on image classification accuracy.

There are other researches applying different kinds of local descriptions for image classification. Ahonen *et al.* [32] analyze the local-binary-pattern (*LBP*) histogram for human face recognition. Similarly, Jabid *et al.* [33] explore the local-directional-pattern (*LDP*) histogram for human face recognition. Xiao *et al*. [34] evaluate numerous scene classification methods using their Sun database. Burl *et al.* [11] propose the constellation model to describe the global geometry of the local observations. Fergus *et al.* [35] combine salient region (*SR*) detection [36] and probabilistic distribution learning to model the global geometry of the local observations. They successfully detect multiple objects by using part-based constellation model. Li *et al.* [2] improve *SR* detection [36] and propose Bayesian decision for image categorization. However, their method cannot detect mirror images and deformable objects. Lin *et al.* [37] develop a parallel averaging stochastic gradient descent (ASGD) algorithm for training one-against-all SVM classifiers. Yu *et al.* [38] proposes an adaptive hypergraph learning method for image classification by generating hyper-edges by linking images and their nearest neighbors.

Ciresan *et al*. [39] propose artificial neural network architectures with minimal receptive fields of convolutional winner-take-all neurons yield large network depth. Recently, deep learning has been proposed for image classification. Zhong *et al*. [40] propose a deep learning model called bilinear deep belief network (*BDBN*) for image classi-

fication. The convolutional neural networks (CNN) [23, 41, 42] have been proposed for large-scale image recognition which has become possible due to the large public image repositories, such as ImageNet [7]. The ConvNets requires high-performance computing systems, such as GPUs or large-scale distributed clusters. Simonyan *et al.* [44] make a thorough evaluation of ConvNets of increasing depth using an architecture with very small convolution filters.

In particular, an important role in the advance of deep visual recognition architectures has been played by the ILSVRC [43], which has served as a testbed for large-scale image classification systems. Image classification has become a great issue in the computer vision, a number of attempts have been made to improve the original architecture in a bid to ILSVRC for achieving better recognition accuracy. For instance, Krizhevsky *et al*. [23] proposing deep ConvNets were the winner of ILSVRC-2012, and Simonyan *et al.* [44] secured the first place in ILSVRC-2014, respectively.

Here, we propose an unsupervised scale-invariant learning to model the regions of interest (*ROI*) based on the probability distribution of the appearance of keypoint (*KP*) and *SR*. Based on the local *ROI* observations, we develop a so-called a *Pairwise Local Observation-based Naive Bayes* (*NBPLO*) classifier for image classification. Our system consists of *KP* extraction, *SR* detection, feature vectors transformation, pairwise *SRs*, and regression model training for each class by using random forest. The local observations are described by *BoF* descriptor. For each input image, we compute the class probability of each local observation pair using regression model and calculate the likelihood of class for the overall observations using naïve Bayes assumption. Finally, we categorize the input image using the maximum likelihood estimation (*MLE*).

2. SYSTEM FRAMEWORK

The proposed image classification system relies on multiple local observations. In each image, we find the *salient regions* (*SR*s) by using Kadir's method [36] to represent the image object. Then, we extract the keypoints (*KP*s) which are used to find the meaningful *SR*s as *feature SR*s. We apply dense *SIFT* [13] to convert each feature *SR* to a multi-dimensional feature vector, and use multiple local pairwise observations to describe the image visual content. The semantically related feature *SR*s are selected as the multiple local pairwise observations which can be used to interpret the image visual content more accurately than the conventional image feature vector model. We assume that every two neighboring feature *SR*s are somehow related. The likelihood of image category is determined by multiple independent pairwise adjacent observations. As shown in Fig. 1 (a), the circles show the feature *SR*s, and the neighboring feature *SR*s are connected, the black line segments connect *SR*s of the same scale, and the red line segments connect *SR*s of different scales.

To find the invariant description to encode the local observations, we apply scale invariant *KP* detection [9] to find discriminative *SIFT* characteristics of the image based on the scale-normalized Laplacian of Gaussian (*LoG*). Then, we eliminate the densely populated *KP*s and create the local image descriptor based on the gradient of the region around the *KP*. Finally, we use the descriptor around *KP* for matching and recognition.

Here, we use *SURF* [10] to find the *KP*s which is a scale and rotation invariant fea-

ture descriptor. Different from Harris corner detection [11] and *SIFT* [9], the *SURF* descriptor is based on Hessian matrix and the distribution of Haar-wavelet responses within the neighborhood of a *KP*. To generate *SURF*, we construct a square region centered around the designated point and along the selected orientation. The region is then split into smaller 4×4 sub-regions. The Haar wavelet responses in horizontal and vertical directions are accumulated for each sub-region as

(a) (b)

Fig. 1. (a) An example image with *SR*s; (b) Adjacent relationship of which the black line segments connect the *SR*s of similar scales and the red line segments connect *SR*s of different scales.

$$\boldsymbol{v} = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|) \tag{1}$$

where d_x and d_y corresponds to Haar wavelet responses in horizontal and vertical direction respectively. The dimension of $\boldsymbol{v}$ is 64 because it contains 16 sub-regions with size 2×2.

We use Bag-of-features (*BOFs*) to represent the *SR*s. The number of *BoFs* (or visual words) is critical for the classification task. The optimal vocabulary size (or the number of *BOFs*) depends on the selected database and the classification model [45]. Similar conclusion [46] was made based on TRECVID and PASCAL database. We apply online spherical *k*-means (*OSKM*) [47] using the Winner-Take-All competitive learning. The learning rate of *OSKM* is exponentially proportionally to the data cluster size.

Then, we propose the *BoF* descriptor, a feature vector, of which the dimension is the vocabulary size. The *BoF* descriptor is similar to the sparse coding algorithm of which the words are atoms and the vocabulary is a dictionary. It is a special case of sparse coding when the sparsity is 1. The *BoF* descriptor requires more representative features with different weightings. Here, we adopt the soft-weighting method [17] as

$$t_k = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(\boldsymbol{v}_j, \boldsymbol{\mu}_k) \tag{2}$$

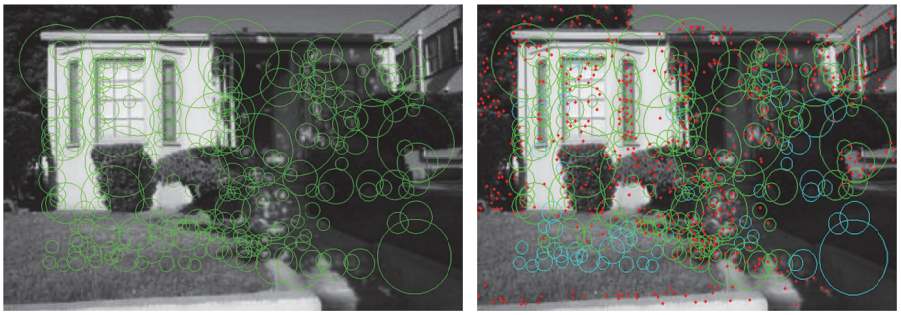
where t_k represents the *k*th component of the *BoF* descriptor T which is a histogram describing the region around *KP i.e.*, $T = [t_1, \ldots, t_K]$, K is the number of *BOFs*, $sim(\boldsymbol{v}_j, \boldsymbol{\mu}_k)$ denotes the similarity between feature vector $\boldsymbol{v}_j$ of the *j*th *KP* and visual word $\boldsymbol{\mu}_k$, $\boldsymbol{\mu}_k$ is the mean vector of the set of feature vectors $\{\boldsymbol{v}_i\}$ assigned to the *k*th group, N is the number of the neighbors, and M_i indicates the number of *KP*s of which the *i*th nearest neighbor is

the visual word μ_k. The soft-weighting of *BoF* assignment outperforms the conventional *BoF* method [25]. It illustrates a significant impact using the soft-weighting scheme which is more robust than the *k-means*.

The local observations are the *SR*s of different scales. We apply the clustering algorithm to find the representative *SR*s as shown in Fig. 2 (a). These *SR*s are local patches located inside the region of interest (*ROI*). Similar to [35], the dimension of *SR*s can be reduced by using *PCA* or *ICA*. The *feature SRs* are found close to the *KPs* as

$$d(\boldsymbol{x}_{KP}, \boldsymbol{x}_{SR}) \leq R_{SR} \tag{3}$$

where $\boldsymbol{x}_{KP}$ is the location of *KP*, d is distance measure, $\boldsymbol{x}_{SR}$ is the center of *SR*, and R_{SR} is the radius of *SR*. The *feature SR* is encoded by *BoF* histogram with soft-weighting.



(a) (b)

Fig. 2. (a) The SRs; (b) The red points correspond to *KP*s; The green circles are the *feature SR*s, whereas the blue ones are the invalid *SR*s.

3. NAIVE BAYES CLASSIFIER

Similar to [2, 35], we propose the structural object class model based on the local observations and their relationship in the images, *i.e.*, $P(\boldsymbol{A}, \boldsymbol{X}|object) \cong P(\boldsymbol{A}, \boldsymbol{X}|\theta_c)$, where θ_c is the object class model of class c, $\boldsymbol{A}$ is the set of the local appearances of object, $\boldsymbol{X}$ is the global structure which is modeled by the local appearances $\boldsymbol{A}$. All possible variations of global structures $\boldsymbol{X}$ can be modeled as the multiple neighborhood relationships of $\boldsymbol{A}$. It is invariant to rotation, translation, and deformation of global structures $\boldsymbol{X}$.

3.1 Local Pairwise Observation

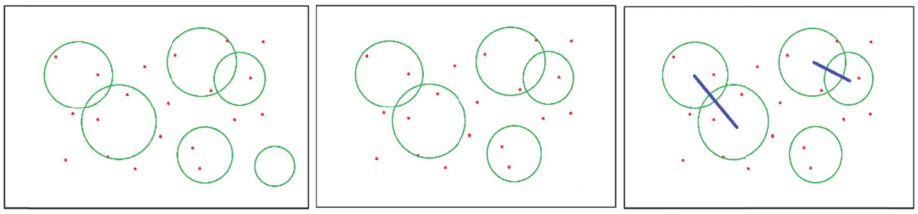
The likelihood of the neighboring local observations of an object is defined as

$$P(A_j \text{ is near to } A_i | A_i, A_j, \theta_c) = P(\delta(A_i, A_j) | A_i, A_j, \theta_c), \tag{4}$$

where θ_c is an object class model of class c, the ith *SR* is denoted as $A_i \equiv SR_i$, If the pairwise local observations A_i and A_j are neighbor then $\delta(A_i, A_j) = 1$, else $\delta(A_i, A_j) = 0$. The neighboring criterion is determined by the geometric relationship of two *SRs* as

$$d(x_{center\ of\ SR_i}, x_{center\ of\ SR_j}) \leq \beta(R_{SR_i} + R_{SR_j}), \tag{5}$$

where SR_i and SR_j are two neighboring *SR*s, β is a relaxation factor with $\beta \geq 1$. The pairwise local observations *SR*s are demonstrated in Fig. 3.



(a) (b) (c)

Fig. 3. The adjacent local observations finding: (a) the detected *KP*s and *SR*s; (b) the *SR*s containing sufficient *KP*s; and (c) the neighboring *SR* pair.

3.2 Regression Object Class Model Training

To generate the object class model $\boldsymbol{\theta}$ of class c_θ based on the pairwise local observations (SR_i and SR_j), we propose two labeling approaches as

Method 1: For bidirectional pairwise local observation (BPLO) descriptor

$$f_{BPLO_{i,j}} = [h_{SR_i}, h_{SR_j}] \text{ is labeled as } l_{BPLO_{i,j}} = \begin{cases} 1 \text{ if } c_f = c_\theta \\ 0 \text{ if } c_f \neq c_\theta \end{cases}.$$

Method 2: For comparative pairwise local observation (CPLO) descriptor

$$f_{CPLO_{i,j}} = [|h_{SR_i}, h_{SR_j}|, h_{SR_i} + h_{SR_j}] \text{ is labeled as } l_{CPLO_{i,j}} = \begin{cases} 1 \text{ if } c_f = c_\theta \\ 0 \text{ if } c_f \neq c_\theta \end{cases}.$$

where h_{SRi}, is the *BoF* histogram of the *i*th *SR*, c_f denotes the class label of local observation pair (*i.e.*, $f_{BPLOi,j}$ or $f_{CPLOi,j}$), and c_θ is the class label of the object class model $\boldsymbol{\theta}$. If the class of the pairwise local observations c_f is the same as the ground truth c_θ, then it is labeled as 1, otherwise, labeled as 0. In our experiments, the first method demonstrates slightly better accuracy than the second one with less memory requirement.

We regard all of the descriptors of the same class as inliers labeled by 1, otherwise as outliers labeled by 0. The outliers which do not belong to the target object are evenly distributed in the images. We apply random forests algorithm [54] to develop the object class model. In random forest training, there are two parameters to be determined: the number of random decision trees and the variables to be determined at each branch node. Different from the other training algorithm, it is non-sensitive to the outliers.

3.3 Naïve Bayes Assumption for Object Recognition

The object class model $\boldsymbol{\theta}_c$ is based on global structure $\boldsymbol{X}$ of observations $\boldsymbol{A}$ as $P(\boldsymbol{A}, \boldsymbol{X}$

$|\boldsymbol{\theta}_c)$ which can also be described as

$$P(\boldsymbol{A}, \boldsymbol{X}|\boldsymbol{\theta}_c) = P(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\theta}_c) \times P(\boldsymbol{A}|\boldsymbol{\theta}_c). \quad (6)$$

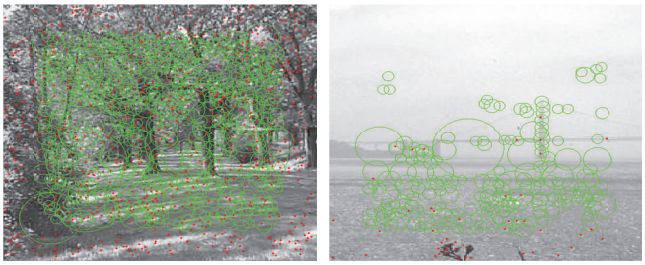
We rewrite the global shape $P(\boldsymbol{X}|\boldsymbol{A}, \theta_c)$ by independent pairwise local observations as

$$P(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\theta}_c) = \prod_{\{i,j\}\in\Psi} P(\delta(A_{SR_i}, A_{SR_j})|A_{SR_i}, A_{SR_j}, \boldsymbol{\theta}_c). \quad (7)$$

where Ψ is the set of local observation pairs. Under the naïve Bayes assumption, we convert the likelihood of the global observation **A** to the likelihood of the set of independent adjacent local observation pairs $\{(A_{SR_i}, A_{SR_j})\}$. The *naïve Bayes classifier algorithm* based on local observation pairs is denoted as Naïve Bayes classifier algorithm using pairwise local observation (*NBPLO*) which is applied to determine the category of the input image with global structure **X** as $c^* = \text{Argmax}_c P(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\theta}_c)$.

4. IMAGE CLASSIFICATION

Here, we show our image classification algorithm which demonstrates better classification accuracy. The *KP*s are related to *SR*s. The *SR*s containing sufficient *KP*s are useful local observations. Fig. 4 shows the images with densely distributed *KP*s and sparsely distributed *KP*s. We may describe *SR* using dense-*SIFT* for comparison with the original *SIFT*-based *BoF* algorithm and the apply pyramid histogram of visual words (*PHOW*) descriptor [32].



(a) (b)

Fig. 4. Example images of (a) densely distributed *KP*s, and (b) sparsely distributed *KP*s.

4.1 Bag of Features with Kernel Weighting

Some *KP*s are not inside but close to the *SR*s. However, these *KP*s do contain information of adjacent *SR*s. We consider the out-of-boundary *KP* features to generate the *BoF* histogram. To avoid the over-influence from the peripheral *KP*s, we add different weights to the *BoF* histogram bins based on the distance from *KP* to the center of the *SR*. The kernel weighting function is a Gaussian distribution:

$$\omega(d, s, \sigma) = e^{\frac{-d^2}{\sigma \cdot s_{SR}^2}} = e^{\frac{-\|x_{SRcenter} - x_{KP}\|^2}{\sigma \cdot s_{SR}^2}}, \quad (8)$$

where d is a 2D distance from the *KP* to the *SR* center, s is the *SR* scale, and σ is an adjustable decay factor.

To determine Gaussian distribution for different decay factors, we choose $\omega \cong 0.7$ or $\sigma = 3$, $d = s$, and $\omega \cong 0.25$ for $\sigma = 3$, $d = 2s$. From Eq. (2), we have *BoF* histogram of the ith *KP* defined as $T_i = [t_1, \ldots, t_K]$. Then, the *BoF* histogram for the ith *SR* is determined based on the soft-weighting assignment to the *BoF* histograms of *M KPs* (*i.e.*, $T_1, T_2, \ldots, T_M$) and described as

$$h_{SR_i} = \frac{\omega_{i,1} T_1 + \omega_{i,2} T_2 + \ldots + \omega_{i,M} T_M}{\omega_{i,1} + \omega_{i,2} + \ldots + \omega_{i,M}} \tag{9}$$

where T_i is the *BoF* histogram describing the region around the ith *KP*, M is the number of *KP*s in the ith *SR*, $\omega_{i,m}$ is the weighting factor of the mth *KP* for the ith *SR*.

4.2 Adjacent Local Observation of Different Scale

The image patches of different scales provide different implications. We may consider two adjacent *SR*s with two significantly different scales. Instead of training one single object class model for all pairwise local observations, we train two object class models: one for handling the pairwise observations with similar scales, and the other for different scales. Here, we split the training dataset into two subsets. One subset consists of similar scale *SR*s, whereas the other subset consists of apparently different scale *SR*s. After training, two object class models $\theta_c^{similar}$ and $\theta_c^{different}$ can be used to find the likelihood of the input image $\boldsymbol{X}$ as

$$\begin{aligned} &P(\boldsymbol{X} \mid \boldsymbol{A}, \boldsymbol{\theta_c}) = P(\delta(A_{SR_i}, A_{SR_j}) \mid A_{SR_i}, A_{SR_j}, \theta_c^{similar}), \\ &\text{and } P(\boldsymbol{X} \mid \boldsymbol{A}, \boldsymbol{\theta_c}) = P(\delta(A_{SR_i}, A_{SR_j}) \mid A_{SR_i}, A_{SR_j}, \theta_c^{different}). \end{aligned} \tag{10}$$

Usually, the detected *SR*s are smaller than the *ROI* around *KP*s. Training the object class model based on these neighboring small *SR*s is inefficient. For the small *SR*s around the same *KP*, their h_{SR_i} are similar. For highly populated *KP*s, these adjacent *SR*s are not effective for object class modeling. We discard the small *SR*s which are close to one another. However, for small object images, we still need to keep small *SR*s.

4.3 Parameters of BoF Soft-Weighting Assignment

The number of neighbors is empirically determined as $N = 4$ [17]. Usually, $N = 3 \sim 6$ shows good experimental results. The *BoF* soft-weighting assignment is defined in Eq. (2). The $sim(\cdot)$ and K are related to N. To compute $sim(\cdot)$, we use inner-product of normalized *KP* features and *SURF* features with the threshold 0.8-0.85. To determine N, we suggest two methods.

Method-1:

$$N \cong \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \delta(sim(\boldsymbol{v}_j, \boldsymbol{\mu}_k) \geq \xi). \tag{11}$$

Method-2: For $N = 1$ to K,

$$\text{Determine } N \text{ if } \xi \cong \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=K-N+1}^{K} sim(\boldsymbol{v}_j, \boldsymbol{\mu}_k), \tag{12}$$

where ξ is a similarity threshold, I is the total number of training images, J is the number of *KP*s, and K is the vocabulary size. These two methods are similar. However, the former is faster than the latter due to less neighbors are chosen so that method-1 is selected.

4.4 Normalization

The defect of naïve Bayes assumption may not be robust for the non-related ingredients, *i.e.*, the novel local observations. Since the non-related ingredient belongs to no category, the class likelihood will deteriorate. For instance, if the probability of an non-related pairwise observation ψ_i is predicted by object class model $P(\theta_{c_1}|\psi_1) = 10^{-3}$ and $P(\theta_{c_2}|\psi_2) = 10^{-7}$. Although the absolute probabilities are both small, $P(\theta_{c_1}|\psi_1)$ is much ($\sim 10^4$ times) larger than $P(\theta_{c_2}|\psi_2)$. To make the object class model more robust to non-related ingredients, we apply *normalization* to remove the relative difference. For each pairwise local observation ψ_i, given a constant λ we calculate the probability of $P(\theta_c|\psi_i)$ for every class object class model. Then, we normalize the probability as $P(\theta_c|\psi_i) \leftarrow \frac{P(\theta_c|\psi_i)+\lambda}{\sum_c [P(\theta_c|\psi_i)+\lambda]}$.

4.5 Training and Testing Process

The preprocessing for training and testing processes consists of detecting condense *KP* features, collecting *KP* feature vectors, using OSKM *algorithm* to find the representative descriptions, and finally determine the number of neighbors N.

Two-class Object class model Training:
Denotations: (1) Each object class model $\boldsymbol{\theta}_c$ consists of $\boldsymbol{\theta}_c^{similar}$ and $\boldsymbol{\theta}_c^{different}$; (2) β is the scale factor of the neighboring *SR*s; (3) σ is the decay factor of Gaussian-distributed weighting kernel; (4) χ is the threshold for determining similar scale adjacent *SR*s; (5) c_f is the true class of local observation pair and c_θ is the class of the object class model $\boldsymbol{\theta}_c$; (6) Each input image consists of two sets of local observation pairs as $\Psi^{similar}$ and $\Psi^{different}$.

For all the training images of class $c = 1, \ldots, C$
(i) Detect the *SR*s.
(ii) Describe each *SR* as h_{SR_i} using Eqs. (8) and (9).
(iii) Find the local observation pair (SR_i, SR_j) satisfying Eq. (5).
(iv) For each local observation pair, if $\chi^{-1} \le R_{SR_i}/R_{SR_j} \le \chi$, then they are similar-scale local observation pairs described as $f_{BPLO_{i,j}}^{similar}$, otherwise as $f_{BPLO_{i,j}}^{different}$.
(v) The local observation pair ($f_{BPLO_{i,j}}^{similar}$ or $f_{BPLO_{i,j}}^{different}$) is labeled as

$$l_{BPLO_{i,j}} = \begin{cases} 1 \text{ if } c_f = c_\theta \\ 0 \text{ if } c_f \neq c_\theta \end{cases}.$$

(vi) Train the object class model $\theta_c^{similar}$ with the training sample set $\{f_{BPLO_{i,j}}^{similar}\}$ using random forest for model regression.
(vii) Train the object class model $\theta_c^{different}$ with the training sample set $\{f_{BPLO_{i,j}}^{different}\}$ using random forest for model regression.

Testing:

For each input image and object class model ($\boldsymbol{\theta}_c^{similar}$, $\boldsymbol{\theta}_c^{different}$)
(i) Detect the *SR*s.
(ii) Describe each *SR* as h_{SR_i} using Eqs. (8) and (9).
(iii) Find the local observation pair (SR_i, SR_j) satisfying Eq. (5).
(iv) If $\chi^{-1} \le R_{SR_i}/R_{SR_j} \le \chi$, then then they are similar-scale local observation pair denoted as $f_{BPLO_{i,j}}^{similar}$, otherwise as $f_{BPLO_{i,j}}^{different}$.
(v) Find $P(\delta(A_{SR_i}, A_{SR_j}) \mid A_{SR_i}, A_{SR_j}, \theta_c^{similar})$.
(vi) Find $P(\delta(A_{SR_i}, A_{SR_j}) \mid A_{SR_i}, A_{SR_j}, \theta_c^{different})$.
(vii) Calculated $P(\boldsymbol{X} \mid \boldsymbol{A}, \boldsymbol{\theta}_c)$ based on the probabilities of all pairwise local observations

$$P(X \mid A, \theta_c) = \sum_{\{i,j\} \in \Psi^{similar}} \ln(P(\delta(A_{SR_i}, A_{SR_j}) \mid A_{SR_i}, A_{SR_j}, \theta_c^{similar})) + \sum_{\{i,j\} \in \Psi^{different}} \ln(P(\delta(A_{SR_i}, A_{SR_j}) \mid A_{SR_i}, A_{SR_j}, \theta_c^{different})).$$

(viii) Categorize the testing image to class c^* based on $c^* = \text{Argmax}_c\, P(\boldsymbol{X} \mid \boldsymbol{A}, \boldsymbol{\theta}_c)$.

5. EXPERIMENTAL RESULTS

In the experiment, we test our system using Scene-15 [4] and Caltech101 [2]. First, we show the confusion matrix of classification results for Scene15 dataset and every 20 categories of Caltech101. The main concern for random forests implementation is the memory space. The memory space limitation disallows testing the entire dataset simultaneously. So, we split Caltech101 dataset into 5 partitions for our experiments. Our experimental results are satisfactory comparing to standard *BoF* method and some classic generative models. Second, we show some examples of probability prediction using neighboring relationship of observations to demonstrate *NBPLO*.

5.1 Scene-15 Dataset

Scene-15 database contains 15 different type of scene, with thirteen classes from [4] (eight classes originate from [5]) and two other classes collected by [4]. Each category has 200 to 400 images, and average image size is 300×250 pixels. Using the same setting as [1, 4], we randomly choose 100 images per class as the training data, and the others as the testing data, with the vocabulary size = 400.

We compare the performance of our method with the other methods [1, 4, 20]. Li's method [4] using the Latent Dirichlet Allocation (*LDA*) is an unsupervised dimensionality reduction technology. It does not achieve high classification accuracy. Labeznik's method [1] partitions the images into increasing fine sub-regions and compute histograms of image feature over the resulting sub-regions. Each image can be described by

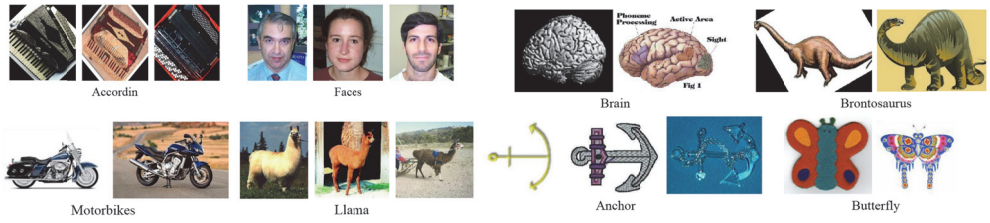
sub-regions in terms of L level and M channels. It generates high dimensional feature vector which can be applied to the so-called Spatial Pyramid Matching (*SPM*) for scene classification. However, *SPM* is very time consuming in high dimensional feature space. The dimensionality of the of the resulting feature vector is $M \cdot (4^{L+1} - 1)/3$. The performance of [1] using so-called the strong features for $L = 0$ and $M = 200$ (or standard Bag of Features) results to the average accuracy of 72.2%. If they reduce the features from $M = 200$ to $M = 60$, the classification rate drops to 63.3% from 72.2%. Table 2 shows that the average accuracy rate of our method (69.68%) is better than standard *BoF* methods [20] (50.55%), and Li's method [4] (65.2%).

Table 1. The average recognition comparisons of using Scene-15 database.

Methods	[1](M=60)	[1](M=200)	[4]	[19]	Ours(NBPLO)
Accuracy rate	63.3	72.2	65.2	50.55	69.68

5.2 Caltech-101 Dataset

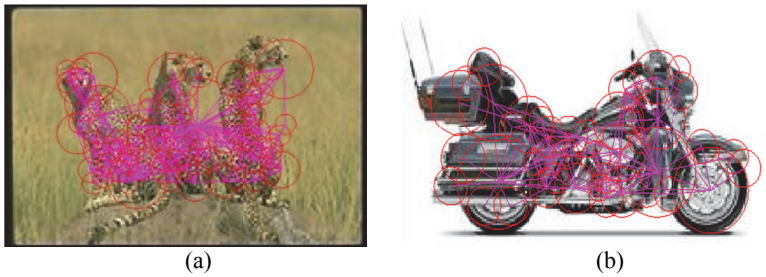
Caltech101 database [2] contains 31-800 images per class of real-world photos and man-made photos as shown in Fig. 5. Our experiments show that our method tends to have better learning capability for the training data from real-world photos. The SIFT feature is mainly for describing local appearances of natural objects. We do not use the category of "BACKGROUND_Google" and "Faces" since the images of the 1st class cannot be well-trained, and the 2nd class is similar to the class "Faces_easy". In our experiments, we take 30 images per class for training, and the others for testing. Similar to previous methods, we quantize the *KP* features into 400 words.



(a) (b)

Fig. 5. Example images for (a) real-world photos and (b) artificial pictures.

Next, some correct classification examples based on the pairwise local observations are shown in Fig. 6. If more than half of local pairwise observations are correctly classified, then the image classification will be correct. The class ambiguity happens because of similar-scale pairwise local observations. It is important to train a model based on different-scale pairwise local observations. As shown in Fig. 6 (b), few distinctive pairwise local observations are located on the wheels. However, it is correctly classified because of the global structure of local observations which models the motorcycle based on the distinctive local observation located on the partial wheel or the entire wheel.



(a) (b)

Fig. 6. Correctly classified images.

Different training images of the same class may not strongly consensus with some of their pairwise local observations. It is difficult to find the regression model truthfully representing the training samples in feature space. Under naïve-Bayes-based assumption, we cannot prevent neglecting the repeated patterns related to the testing images. Our method does not perform the matching process, nor check the completeness or overall appearance. Therefore, the input image categorization is easily affected by repeated but meaningless patterns. Here, we choose the classes with low classification accuracy and high classification accuracy for further testing. Each class contains more than 30 testing images for more accurate verification.

The average accuracy is 47.12%, which is a little better than the result of *BoF* method with accuracy 30.03%. The reason may be the diverge appearance of the training dataset. The local observations do not focus on some specific features in the *BoF* histogram feature space, so that the prediction may fail. The other concern is the shape difference of the objects in the image. The *SR*s capture different information from different shapes and generate incoherence pairwise local observations. The average accuracy of the result is 89.65%, which outperforms the results of *BoF* method (58.24%). If the variation of local observations of the training data is limited, then the improvement will be significant. The comparisons with the other methods are shown in Table 2. Our method has two advantages: (1) random forest for likelihood prediction provides a good regression surface, and (2) the connectivity of pairwise local observations provide discriminative features.

Table 2. Recognition results comparison of using the Caltech 101 database.

Training image	5	10	15	20	25	30
Lazebnik [1]	N/A	N/A	56.40	N/A	N/A	64.60
Griffn [8]	44.2	54.5	59.0	63.3	65.8	67.6
R. Forests [53]	N/A	N/A	70.4	N/A	N/A	80
R. Ferns [53]	N/A	N/A	70	N/A	N/A	79.2
Zhang [49]	46.6	55.8	59.1	62.0	N/A	66.20
Wang [25]	51.15	59.77	65.43	67.74	70.16	73.44
Yang [24]	N/A	N/A	67.0	N/A	N/A	73.2
Gemert [48]	N/A	N/A	N/A	N/A	N/A	64.16
SRC [50]	48.8	60.1	64.9	67.7	69.2	70.7
K-SVD [51]	49.8	59.8	65.2	68.7	71.0	73.2
D-KSVD [52]	49.6	59.5	65.1	68.6	71.1	73.0
LC-KSVD [26]	54.0	63.1	67.7	70.5	72.3	73.6
Ours(NBPLO)	N/A	N/A	N/A	N/A	N/A	85.69

5.3 Demonstrations

Then, we show some correctly classified images containing some dominant objects. In Fig. 7 (a), the *SR*s contain partial appearance of the "*tall buildings*". Figs. 7 (b) and (c) show that most of the pairwise observations are classified to the "*office*" or "*bedroom*". In Fig. 7 (d), observations from most of the products on the shelves belong to the "*store*", while ambiguous observation pairs are at the lamp and the windows. In Fig. 7 (e), the observation pairs of ocean and beach sand are classified to the "*beach*", whereas the observation pairs of the trees, sky, and clouds are ambiguous. In Fig. 7 (f), pairwise observations of trees are classified to the "*forests*", and the ingredients of the "*house*" are ambiguous. In Fig. 7 (e), the ambiguous pairwise observations may apparently be classified to the "*mountain*", however the majority pairwise observations are categorized to the "*beach*". In Fig. 7 (h), the ambiguous pairwise observations are distributed at the ground, while the observations that capture partial appearance of "*trees*" are correctly classified as "*woods*".

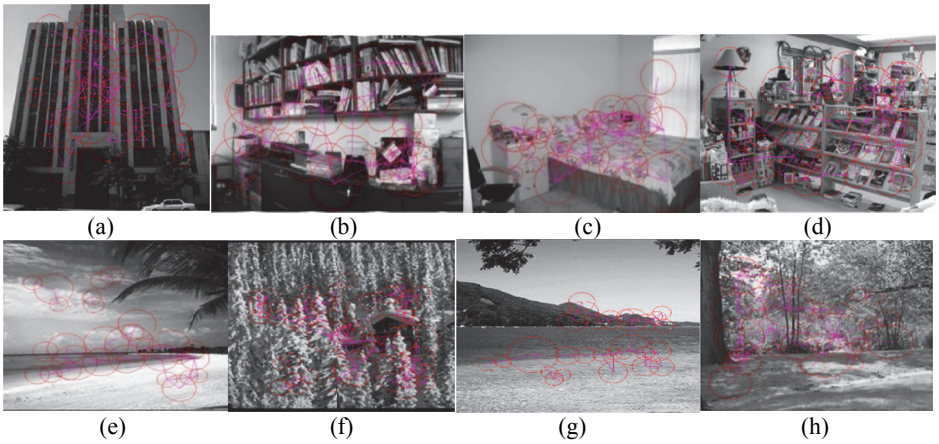


Fig. 7. The results of correct classification.

The experimental results show that our method is effective. However, it may fail for classifying the image containing mixed objects that belong to different classes, such as image of a house in the forest or a beach with large mountains as the background.

6. CONCLUSIONS

We have proposed an image classification method based on local pairwise observations which outperforms previous methods. However, our method still need some improvement. First, we need to relax our naïve-Bayes assumption, which is too "naive" for general image recognition. Second, we have memory space problem, which prevent us from doing experiments on massive database.

REFERENCES

1. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid

matching for recognizing natural scene categories," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2006, pp. 2169-2178.
2. F.-F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, Vol. 106, 2007, pp. 59-70.
3. J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: large-scale scene recognition from abbey to zoo," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485-3492.
4. F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2005, pp. 524-531.
5. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, Vol. 42, 2001, pp. 145-175.
6. M. Everngham, I. Van Gool, C. K. I. Willams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International Journal on Computer Vision*, Vol. 88, 2010, pp. 303-338.
7. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.
8. G. Griffin, A. Holub, and P. Perona, "Caltec-256 object category dataset," Technical Report UCB/CSD-04-1366, Caltech, 2007.
9. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, Vol. 60, 2004, pp. 91-110.
10. B. Herbert, T. Tuytelaars, and L. van Gool, "SURF: Speeded up robust features," in *Proceedings of European Conference on Computer Vision*, 2006, pp. 404-417.
11. C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, 1988, pp. 147-151.
12. N. Dalal, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, 2005, pp. 886-893.
13. G. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on PAMI*, Vol. 33, 2011, pp. 978-994.
14. L. Wu, S. C. H. Hoi, and N. Yu, "Semantics-preserving bag-of-words models for efficient image annotation," in *Proceedings of the 1st ACM Workshop Large-Scale Multimedia Retrieval and Mining*, 2009, pp. 19-26.
15. P. Tirilly, V. Claveau, and P. Gros, "Language modeling for bag-of-visual words image categorization," in *Proceedings of International Conference on Content-Based Image and Video Retrieval*, 2008, pp. 249-258.
16. Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian, "Toward a higher-level visual representation for object-based image retrieval," *Visual Computer*, Vol. 25, 2008, pp. 13-23.
17. Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, 2007, pp. 494-501.
18. K. Kesorn and S. Poslad, "An enhanced bag-of-visual word vector space model to

represent visual content in athletics images," *IEEE Transactions on Multimedia*, Vol. 14, 2012, pp. 211-222.
19. J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, Vol. 105, 2013, pp. 222-245.
20. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1-2.
21. J. Sivic and A. Zisserman, "Video google: "A text retrieval approach to object matching in videos," in *Proceedings of International Conference on Computer Vision*, 2003, pp. 1470-1477.
22. K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *Journal of Machine Learning Research*, Vol. 8, 2007, pp. 725-760.
23. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of Conference on Neural Information Processing Systems*, 2012, pp. 1097-1105.
24. J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *IEEE Computer Vision and Pattern Recognition*, 2009, pp. 1794-1801.
25. J. Wang, J. Yang, K.Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," *IEEE Computer Vision and Pattern Recognition* 2010, pp. 3360-3367.
26. Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Transactions on PAMI*, Vol. 35, 2013, pp. 2651-2664.
27. J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, "What is the spatial extent of an object?" in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 770-777.
28. E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proceedings of European Conference on Computer Vision*, 2006, pp. 490-503.
29. F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proceedings of International Conference on Computer Vision*, 2005, pp. 604-610.
30. J. Zhang, M. Marszalck, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal on Computer Vision*, Vol. 73, 2006, pp. 213-238.
31. Y. Jia, C. Huang, and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3370-3377.
32. T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on PAMI*, Vol. 28, 2006, pp. 2037-2041.
33. T. Jabid, M. H. Kabir, and O. Chae, "Local directional pattern (LDP) for face recognition," in *Proceedings of International Conference on Consumer Electronic*, 2010, pp. 329-330.
34. M. C. Burl, M. Weber, and P. Perona, "A probabilistic approach to object recogni-

tion using local photometry and global geometry," in *Proceedings of European Conference on Computer Vision*, 1998, pp. 628-641.
35. R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2003, pp. 264-271.
36. T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal on Computer Vision*, Vol. 45, 2001, pp. 83-105.
37. Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour and K. Yu, L. Cao and T. Huang, "Large-scale image classification: Fast feature extraction and SVM training," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1689-1696.
38. J. Yu, D. Tao, and M. Wang, "Adaptive hypergraph learning and its application in image classification," *IEEE Transactions on Image Processing*, Vol. 21, 2012, pp. 3262-3272.
39. D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3262-3272.
40. S.-H. Zhong, Y. Liu, and Y. Liu, "Bilinear deep learning for image classification," in *Proceedings of ACM International Conference on Multimedia*, 2011, pp. 343-452.
41. J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng, "Large scale distributed deep networks," in *Proceedings of Annual Conference on Neural Information Processing Systems*, 2012, pp. 1232-1240.
42. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *Proceedings of International Conference on Learning Representations*, 2014.
43. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li, *Imagenet Large Scale Visual Recognition Challenge*, Vol. 115, 2014, pp. 211-252.
44. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of International Conference on Learning Representations*, 2015.
45. C.-F. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artificial Intelligence*, Vol. 2012, 2012, Article ID 376804.
46. J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of ACM International Workshop on Multimedia Information Retrieval*, 2007, pp. 197-206.
47. S. Zhong, "Efficient online spherical *k*-means clustering," in *Proceedings of IEEE International Joint Conference on Neural Network*, Vol. 5, 2005, pp. 3180-3185.
48. J. Gemert, J. Geusebroek, C. Veenman, and A. Smeulders, "Kernel codebooks for scene categorization," in *Proceedings of European Conference on Computer Vision*, 2008, pp. 696-709.
49. H. Zhang, A. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2126-2136.
50. J. Wright, M. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via

sparse representation," *Transactions on PAMI*, Vol. 31, 2009, pp. 210-227.
51. M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, Vol. 54, 2006, pp. 4311-4322.
52. Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2691-2698.
53. A. Bosch, A. Zisserman, and X. Muoz, "Image classification using random forests and ferns," in *Proceedings of International Conference on Computer Vision*, 2007, pp. 1-8.
54. L. Breiman, "Random forests," *Machine Learning*, Vol. 45, 2001, pp. 5-32.

Shih-Chung Hsu (徐士中) received the BS and MS degrees from National Taiwan University of Science and Technology, Taipei, Taiwan, in 2004 and 2006, respectively. Currently, he is a Ph.D. student in Electrical Engineering Department, National Tsing-Hua University, Hsinchu, Taiwan. His research interests include computer vision, pattern recognition, machine learning.

I-Chieh Chen (陳羿捷) received the BS degree and MS degree in Electrical Engineering from National Tsing-Hua University, Hsin-Chu, Taiwan, in June 2012 and September 2014, respectively. His research interests include computer vision, pattern recognition, machine learning. His current position is an Engineer in Micron Inc., Taichung, Taiwan.

Chung-Lin Huang (黃仲陵) received his BS degree from the National Tsing-Hua University and MS degree from National Taiwan University respectively. He obtained his Ph.D. degree in Electrical Engineering from the University of Florida, Gainesville, FL, USA, in 1987. From 1988 to 2012, he was a Professor in the Electrical Engineering Department, National Tsing-Hua University, Hsinchu, Taiwan. Since August 2012, he has been a Professor in Department of M-Commerce and Multimedia Applications, Asia University, Taichung, Taiwan. He has received numerous paper awards from International and Local Academic Conferences. His research interests are in the area of image processing, computer vision, and visual communication.