Image Classification Using Pairwise Local Observations Based Naive Bayes Classifier

Shih-Chung Hsu[*], *I-Chieh Chen*[*] *and Chung-Lin Huang*[†]
[*]*Electrical Engineering Department, National Tsing-Hua University, Hsin-Chu, Taiwan*
[†]*Department of M-Commerce and Multimedia Applications, Asia University, Tai-Chung, Taiwan*
E-mail: clhuang@asia.edu.tw, Tel: +886-4-23323456

***Abstract*—We present an image classification method which consists of salient region (*SR*) detection, local feature extraction, and pairwise local observations based Naive Bayes classifier (*NBPLO*). Different from previous image classification algorithms, we propose a scale, translation, and rotation invariant image classification algorithm. Based on the discriminative pairwise local observations, we develop the structure object model based Naive Bayes classifier for image classification. We do the experiments using Scene-15 and Caltech-101 database and compare the experiment results of bag-of-features (*BoF*) and *SPM* algorithms.**

***Keywords*—Image classification, Salient Region (*SR*), Bag-of-Features (*BoF*), Pairwise Local Observation Based Naïve Bayes (*NBPLO*) Classifier.**

I. INTRODUCTION

Image classification has been a challenging problem due to the complexity and the variety of images. The image classification methods can be categorized to three different approaches. The first category is the edge-based feature point detection such as Scale Invariant Feature (*SIFT*) [1], Speed-up Robust Feature (*SURF*) [2], and Harris corner detection [3]. The second category is histogram-based image representation such as bag-of-features (*BoF*), and spatial-pyramid-matching (*SPM*) [4]. The third category is coding-based image representation such as sparse coding (*SRC*) [5], Locality-constrained Linear Coding (*LLC*) [6], and Label Consistent K-SVD (*LC-KSVD*) [7].

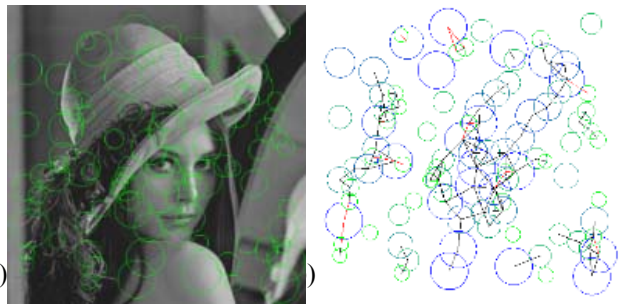
Most of the researches apply histogram-based image descriptions of which the main difference is visual words calculation. Ahonen *et al.* [8] use local-binary-pattern (*LBP*) histogram for human face recognition. Similarly, Jabid *et al.* [9] adopt local-directional-pattern (*LDP*) histogram for human face recognition. Lazebnik *et al* [4] quantize *SIFT* feature and oriented edge points into strong and weak features respectively for pyramid matching. In [4, 10, 16, 33, 28, 29], they do the statistical analysis of visual words assignment for global or local patch and show reasonable results of scene recognition.

Burl *et al.* [11] propose the constellation model to describe the global geometry of the local observations. Fergus *et al.* [12] combine the salient region (*SR*) detection method [13] and the probabilistic distribution learning method to model the global geometry of the local observations. They successfully detect multiple objects by using part-based constellation model. To improve the *SR* detection method, *Li et al.* [14] use Bayesian decision method for image category recognition and detection.

We propose an unsupervised scale-invariant model based on multiple local observations. First, we introduce the *KP* detection and *SR* detection method. Second, we transform the features to visual words by using feature quantization method and *BoF* assignment method. Third, we show how to represent *SR* by multiple *KP*s. Fourth, we develop a method to classify the images based on the pairwise local observations.

We define the *basic part* (*BP*) to represent the appearance of the designated object, and the *salient regions* (*SR*) to represent the region of interest. The *SR* may not be a *BP*, however, the *BP* is a *SR*. Each object is represented by many *BP*s. The priori probability of the existence of any *BP* is a constant. The existence of *BP* is independent of the corresponding object.

Two neighboring *BP*s are related. The joint likelihood of multiple *BP*s can be simplified by multiple pairwise adjacent observations as $\prod_{\{Bi,Bj\}\in\Psi} P(B_i, B_j)$, where $\Psi = \{(Bi, Bj)\}$ is a set of pairwise *BP*s. As shown in Fig. 1, the circles of the same color show the *SR*s of the same scale, the green ones are larger *SR*s, and the brown ones are smaller *SR*s. The connected *SR*s are neighbors. The purple lines connect *SR*s of the same scale, whereas the red lines connect two *SR*s of different scales.



(a) (b)

Fig. 1. (a) an example image with *SR*s. (b) adjacent relationship of similar scale *SR*s and different scale *SR*s.

The preprocessing process consists of feature extraction, feature vector quantization, *SR* detection, *KP* finding, neighboring *SR* finding and description. For each image class, we apply the regression model training by using the random forest method [27]. For each testing image, we use the

regression model to calculate the class likelihood of all local observations based on naïve Bayes assumption. Finally, we can determine the image class by using maximum likelihood estimation (*MLE*).

II. Feature Extraction

Here, we normalize the extracted local patches as the local observations, and find a rotation and translation invariant description method to encode the local observations.

A. SIFT

We apply the scale invariant *KP* detection [1] to find the discriminative *SIFT* characteristics [15, 16] of the image using the difference of Gaussian and the scale-normalized Laplacian of Gaussian (*LoG*). Then, we eliminate the densely populated *KP*s and create the local image descriptor (feature vector) based on the gradient of the region around the *KP*. Finally, we use the descriptor around *KP* for matching and recognition.

B. SURF

SURF [2] is a better feature than *SIFT* as a scale and rotation invariant feature descriptor. Different from Harris corner detection [3] and SIFT [1], SURF descriptor is based on Hessian matrix and a distribution of Haar-wavelet responses within the neighborhood of a *KP*. SURF is 64 dimension feature vector for fast matching and robustness based on Hessian matrix as

$$\mathcal{H}(\boldsymbol{x},\sigma) = \begin{bmatrix} L_{xx}(\boldsymbol{x},\sigma) & L_{xy}(\boldsymbol{x},\sigma) \\ L_{xy}(\boldsymbol{x},\sigma) & L_{yy}(\boldsymbol{x},\sigma) \end{bmatrix} \quad (1)$$

where $L_{xx}(\boldsymbol{x},\sigma)$ is the convolution of the 2nd derivative Gaussian with the image. The descriptor represents a square region centered around the designated point and along the selected orientation. The region is split into smaller 4×4 sub-regions. The Haar wavelet response in horizontal and vertical directions are accumulated for each sub-region which is represented by a four-dimensional descriptor vector $\boldsymbol{v}$ as

$$\boldsymbol{v} = (\Sigma d_x, \Sigma d_y, \Sigma|d_x|, \Sigma|d_y|) \quad (2)$$

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

C. Condense SIFT

Condense features often provide more valid information for the classifier to generate better results compared with the sparse features. Liu, *et al.* [18] compute the top three principal components of *SIFT* descriptors, and then convert the principal components to *RGB* space. We extract the feature description using *SIFT* feature [19, 20].

III. Bag-of-Feature

Bag-of-features (or visual words) are widely used for image recognition due to its robustness to image rotation, scale, and translation. The number of bag-of-features (*BoF*) is crucial for image classification applications. The optimal vocabulary size (or the number of visual words) depends on the image database and the classification model [21]. In [22], they made similar conclusion based on *TRECVID* and *PASCAL* database to determine the vocabulary size

A. Online Spherical K-means Algorithm

To normalize the training dataset, we apply online spherical *k*-means (*OSKM*) [23] based on the Winner-Take-All competitive learning. The learning rate (the number of iterations) for *OSKM* increases exponentially as the data cluster size increases. The on-line learning achieves significantly better clustering results than the batch learning. The *KP* features are required for the *OSKM* algorithm to find the representative descriptions.

B. BoF with Soft-Weighting

We apply *BoF* algorithm to represent the images for categorization, and then use the regression method to model the *BoF* histogram of each category. Each *BoF* histogram is a feature vector of which the dimension *K* is the vocabulary size. To represent the images using *BOF* effectively, we apply the *BoF* soft-weighting method [24] as

$$t_k = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(\boldsymbol{v_j}, \boldsymbol{\mu}_k) \quad (3)$$

where t_k represents the *k*th value of the *K*-D *BoF* histogram vector. *K* indicates the number of visual words. *N* is the number of the neighbors in *k*-nearest-neighbor (*kNN*). M_i indicates the total amount of *KP*s of which the *i*th nearest neighbors is the k^{th} visual words $\boldsymbol{\mu}_k$. The soft-weighting *BoF* assignment outperforms the conventional *BoF* method[24].

We compare condense-*SIFT* with the original *BoF* algorithm using *SIFT* and choose a scale-invariant method, *i.e.*, pyramid histogram of visual words (*PHOW*) descriptor [31].

IV. Pairwise Local Observation Based Naive Bayes Classifier

Here, we propose a structural image object model based on local observations and their relationship [12, 14, 25]. It can be depicted as $P(\boldsymbol{A}, \boldsymbol{X} \mid \theta_c)$, where θ_c is the image object model of class *c*, $\boldsymbol{A}$ is the set of the local appearances, and $\boldsymbol{X}$ is the global structure. The image object can be modeled by multiple neighboring local observations (or *SR*s) invariant to rotation, translation, and deformation.

A. SR Detection

To find useful local observations, we apply *SR* detection [13] based on the Shannon entropy as

$$H_{D,R_X} = -\sum_i P_{D,R_X}(d_i) log_2 P_{D,R_X}(d_i), \quad (4)$$

where $P_{D,R_X}(d_i)$ is the probability of descriptor *D* taking the value d_i in the local region R_X. d_i is the i^{th} bin of image intensity histogram.

The *SR* detection method [13] finds the candidate *SR*s of different scales. There are about 60,000 candidate *SR*s. We apply the clustering algorithm to find the representative *SR*s instead of the *k-means* algorithm due to the cluster number is required as a-priori. *Kadir et al*. [13] developed a simple greedy clustering algorithm to solve the problem. The identified *SRs* are shown in Fig. 2.

Fig. 2. The result of *SR* detection.

B. SR Description

We divide each image into patches as the local observations, normalize the patches, and reduce the dimension of the patches by using *PCA* or *ICA*. However, the local observations (or *SR*s) located outside the region of interest (*ROI*) are unreliable observations. Different from [12], the *KP*s are captured by using the SIFT features as shown in Fig. 3. Then we find the reliable *SRs* which are closed to the KPs as the *feature SRs* based on

$$d(\boldsymbol{x}_{KP}, \boldsymbol{x}_{SR}) \leq R_{SR} \tag{5}$$

where $\boldsymbol{x}_{KP}$ is the location of *KP*, $\boldsymbol{x}_{SR}$ is the center of *SR*, d is distance measure, and R_{SR} is the radius of *SR*. We encode each *feature SR* by using *BoF* with soft-weighting assignment and concatenate the *BoF* histograms from different *KP*s as the local discriminative observation.

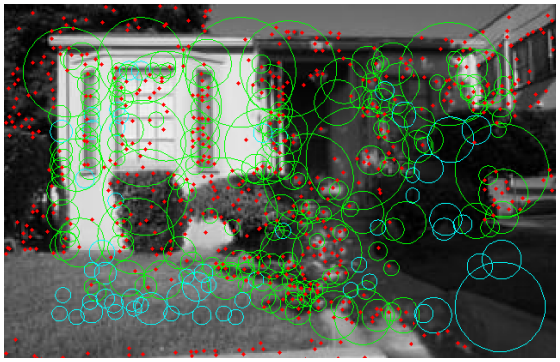


Fig. 3. The key points (*KP*s) in red correspond to the detected SIFT features, whereas the green circles are the feature *SR*s, whereas the blue ones are the neglected *SR*s.

C. Local Pairwise Observation

The likelihood of the neighboring local observations is defined as $P(\delta(A_i, A_j)|A_i, A_j, \theta_c)$ where θ_c is an image object model of class c, $\delta(A_i, A_j)$ indicates the neighboring local observations A_i and A_j. The i^{th} *SR* is a local observation defined as A_i and the geometric relationship between two *SRs* is defined as

$$\delta(A_i, A_j) \leq \beta\left(R_{SR_i} + R_{SR_j}\right), \tag{6}$$

where R_{SR} is the radius of *SR*, β is a relaxation factor with $\beta \geq 1$. The pairwise local observations are demonstrated in Fig. 4.

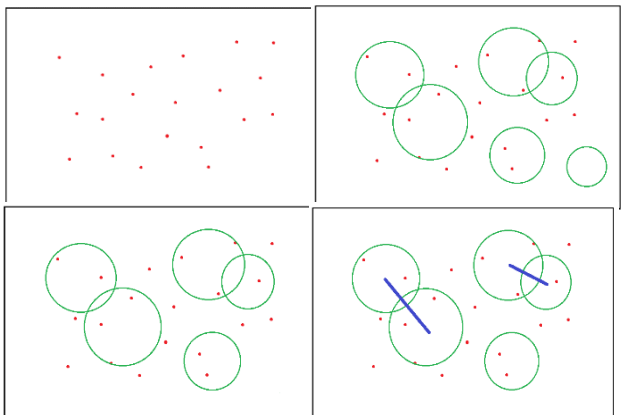


Fig. 4. The adjacent local observations finding consists of (a) locate all *KP*s, (b) the detected *SR*s, (c) the *SRs* containing sufficient *KP*s, and (d) the neighboring *SR* pair.

D. Regression Model Training

To model the local pairwise observations, we propose two approaches.

Method 1. Bidirectional description for pairwise local observations (BPLO)

$$f_{BPLO_{i,j}} = [\, h_{SR_i}, h_{SR_j}\,]; \quad l_{BPLO_{i,j}} = \begin{cases} 1 \; if \; c_f = c_\theta \\ 0 \; if \; c_f \neq c_\theta \end{cases}. \tag{7}$$

Method 2. Comparative description for pairwise local observations (CPLO)

$$f_{CPLO_{i,j}} = [\, |h_{SR_i} - h_{SR_j}|\,,\; h_{SR_i} + h_{SR_j}]; \tag{8}$$

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

where h_{SR_i} is the *BoF* histogram of the i^{th} *SR*, c_f denotes the class of the pair of local observation descriptor (*i.e.*, $f_{BPLO_{i,j}}$ or $f_{CPLO_{i,j}}$), and c_θ is the class of object model θ. If the class of pairwise local observations is identical to the designated object model, then they are labeled as 1, otherwise, labeled as 0. In our experiments, we find that the 1st method demonstrates slightly better accuracy than the 2nd method with less memory.

We regard the local descriptions of the same class as inliers, otherwise as outliers. The outliers do not belong to the target object and uniformly distributed in images. We apply random forests algorithm [27] for the object model training. There are two parameters to be determined: the number of random decision trees and the thresholds to be determined at each branch node. Different from the other training algorithm, it is non-sensitive to the outliers. Our local observations are described by histogram-like feature vectors.

E. Naïve Bayes Assumption for Object Recognition

Similar to [12, 14], the image object model $\boldsymbol{\theta}_c$ is based on global structure $\mathbf{X}$ of observations $\mathbf{A}$ as $P(\boldsymbol{A}, \boldsymbol{X} | \boldsymbol{\theta}_c)$ which can also be described as

$$P(\boldsymbol{A}, \boldsymbol{X} \mid \theta_c) = P(\boldsymbol{X} \mid \boldsymbol{A}, \theta_c) \times P(\boldsymbol{A} \mid \theta_c). \quad (9)$$

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

$$P(\boldsymbol{X} \mid \boldsymbol{A}, \theta_c) = \prod_{\{i,j\} \in \Psi} P\left(\delta(A_{SR_i}, A_{SR_j}) \middle| A_{SR_i}, A_{SR_j}, \theta_c\right). \quad (10)$$

Under the naïve Bayes assumption, we convert the object model overall structure to the likelihood of the independent adjacent local observations. We propose the *naïve Bayes classifier algorithm* based on local pairwise observations.

Pairwise Local Observation Based Naïve Bayes (*NBPLO*) Classifier

Denotation: k is the number of centers for k-means algorithm, β is the scale factor for the neighborhoods relationship of *SR*s; N is the number of neighbors. $f_{BPLO_{i,j}}$ is the pair of local observation descriptor. Collect KPs and use *OSKM* to find the representative descriptions as $SR_1, SR_2, \ldots, SR_n$.

I. Training:

For each class object model and each training image:

i) Use *BoF* soft-weighting assignment to describe the appearance of each *SR* as A_{SR_i}.

ii) Find two adjacent *SR*s if they satisfy the following

$$d\left(x_{center\ of\ SR_i}, x_{center\ of\ SR_j}\right) \leq \beta(R_{SR_i} + R_{SR_j}) . \quad (11)$$

iii) Generate $f_{BPLO_{i,j}}$ as the training data.

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

where c_f denotes the class of the pair of local observation descriptor (*i.e.*, $f_{BPLO_{i,j}}$), and c_θ is the class of object model θ. We train the regression model of object class c as θ_c by using random forest algorithm of all $l_{BPLO_{i,j}}$ corresponding to $f_{BPLO_{i,j}}$.

II. Testing:

For each testing image and each object model of class c as θ_c

i) Use *BoF* soft-weighting assignment to describe the appearance of the i^{th} *SR* as A_{SR_i}.

ii) Determine two adjacent *SR*s using Eq. (11)

iii) Predict the probability $P\left(\delta(A_{SR_i}, A_{SR_j}) \middle| A_{SR_i}, A_{SR_j}, \theta_c\right)$

iv) Calculate

$P(\boldsymbol{X} \mid \boldsymbol{A}, \theta_c) = \sum_{\{i,j\} \in \Psi} ln\ (P(\delta(A_{SR_i}, A_{SR_j}) \mid A_{SR_i}, A_{SR_j}, \theta_c))$

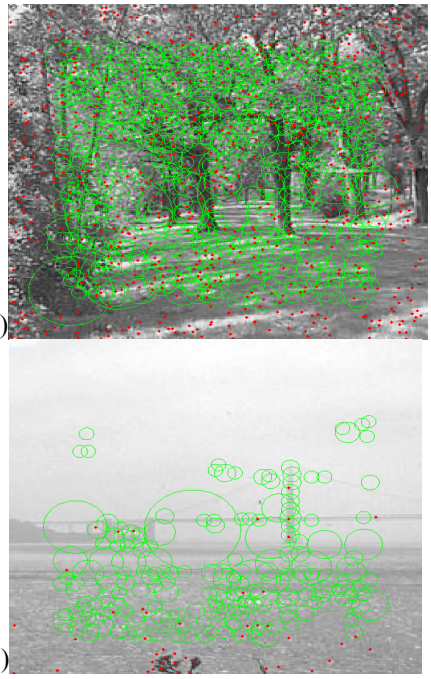
The category of the testing image is determined by using the maximum log-likelihood estimation (*MLE*).

V. Implementation

Here, we implement our image classification algorithm which shows better classification accuracy.

A. The Empty SRs

We need to find the meaningful *SR*s with *KP*s. If *SR* contains no *KP*, then it an empty *SR*. However, the empty *SR*s still provide some information which can be converted to local descriptors. Fig. 5 shows the images with densely distributed *KP*s and sparsely distributed *KP*s. If most of the *SR*s contain no *KP*, then the training process using adjacent pairwise observations will become meaningless.



(a) (b)

Fig. 5. Example images of (a) densely distributed *KP*s, and (b) sparsely distributed *KP*s.

B. BoF with Weighting

However, there are some *KP*s not inside but close to the *SR*s. These *KP*s do contain information extracted from some part of images overlapping with the *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 weight the *BoF* histogram based on the distance from *KP* to the center of the *SR*. The kernel function is a Gaussian distribution:

$$\omega(d, s, \sigma) = e^{\frac{-d^2}{\sigma \cdot s_{SR}^2}} = e^{\frac{-||x_{SR\ center} - x_{KP}||^2}{\sigma \cdot s_{SR}^2}}, \quad (12)$$

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

To determine Gaussian distribution parameters, we choose $\omega \cong 0.7$ for $\sigma = 3, d = s$, and $\omega \cong 0.25$ for $\sigma = 3, d = 2s$. The soft-weighting assignment to *BoF* histograms of *KP* features $T_1, T_2, \ldots, T_M$ is

$$h_{SR_i} = \frac{\omega_{i,1}T_1 + \omega_{i,2}T_2 + \cdots + \omega_{i,M}T_M}{\omega_{i,1} + \omega_{i,2} + \cdots + \omega_{i,M}} \quad (13)$$

where $\omega_{i,m}$ is the weighting factor of the m^{th} *KP* and the i^{th} *SR*, and h_{SR_i} is the soft-weighting *BoF* histogram for the i^{th} *SR*.

C. Adjacent Local Observation with Different Scale

The local patches (or *SR*s) of different scales provide different implications. For all pairwise local observations, we train two object models, $\theta_c^{similar}$ and $\theta_c^{different}$, one for similar scale, and the other for different scales, *i.e.*,

$$P\left(\delta(A_{SR_i}, A_{SR_j}) \middle| A_{SR_i}, A_{SR_j}, \theta_c^{similar}\right)$$

and $P\left(\delta(A_{SR_i}, A_{SR_j}) \middle| A_{SR_i}, A_{SR_j}, \theta_c^{different}\right)$.

The small *SR*s are useless for object model which are discarded. However, for small images, the small *SR*s are still valuable.

D. BoF Soft-Weighting Assignment

The number of neighbors is empirically determined as N=4 [24]. Usually, N=3~6 shows good results for the experiments. The *BoF* soft-weighting assignment is

$$t_k = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{1}{2^{i-1}} sim(\boldsymbol{v_j}, \boldsymbol{\mu_k}), \quad (14)$$

where t_k represents the kth bin of the K-D *BoF* histogram vector, and $sim(\boldsymbol{v_j}, \boldsymbol{\mu_k})$ denotes the similarity between feature vector $\boldsymbol{v_j}$ and visual word $\boldsymbol{\mu_k}$. The number of visual words is K. N denotes the number of the neighbors in *KNN* process. M_i describes the total amount of *KP*s of which the ith nearest neighbors is $\boldsymbol{\mu_k}$. The $sim(\cdot)$ and K are related to N. To compute $sim(\cdot)$, we use inner-product method of normalized *KP* features and SIFT feature with the threshold 0.8~0.85. To determine N, we suggest two methods.

Method-1:

$$N \cong \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\delta(sim(\boldsymbol{v_j}, \boldsymbol{\mu_k}) \geq \xi).$$

Method-2: For N = 1 to K,

Determine N if $\xi \cong \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=K-N+1}^{K} sim(\boldsymbol{v_j}, \boldsymbol{\mu_k})$,

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 equivalent. Here, we choose the first method which is faster than the second one due to less neighbors are chosen.

E. The Influence of Single Observation

We assume that a single *BP* is independent of the observed object. The likelihood of image $\boldsymbol{A}$ is the product of the likelihoods of multiple *BPs as*

$$P(\boldsymbol{A} \mid \theta_c) = \prod_{i=1}^{n} P\left(A_{SR_i} \middle| \theta_c\right) = C^n.$$

So, we modify the likelihood of image $\boldsymbol{A}$ as

$$P(\boldsymbol{A}, \boldsymbol{X} \mid \theta_c) \propto \prod_{i=1}^{n} P\left(A_{SR_i} \middle| \theta_c\right)^{\alpha} \prod_{\{i,j\}\in\Psi} P\left(\delta(A_{SR_i}, A_{SR_j}) \middle| A_{SR_i}, A_{SR_j}, \theta_c\right)^{1-\alpha}$$

where $0 \leq \alpha \leq 1$ is a relative weighting parameter for adjusting the above two terms. After the object model training, we have $P\left(A_{SR_i} \middle| \theta_c\right)$. We determine the best relative weighting parameter by iteratively adjusting α by a step size $\Delta\alpha$=0.1 in each iteration. We use Caltech101 database of 18 classes in our experiments. Taking 30 images per class for training and the others for testing, we find that the average accuracy drops when the relative weighting of the appearance increases.

F. Normalization

The naïve Bayes assumption may not be robust for the non-related ingredients, *i.e.,* the novel local observations. Since the non-related ingredient does not belong to any category, its likelihood will deteriorate to relatively low. For instance, the likelihood of two non-related pairwise observations ψ_1 and ψ_2 are predicted by two object models as $P(\theta_{c_1} \mid \psi_1) = 10^{-3}$ and $P(\theta_{c_2} \mid \psi_2) = 10^{-7}$ which are both small. However, the former is much larger than the latter. To make the object model more robust to non-related ingredients, we apply *normalization* to remove the relative difference. For each pairwise local observation ψ_i, we choose a constant λ to normalize the probability as $P(\theta_c \mid \psi_i) \leftarrow \frac{P(\theta_c \mid \psi_i)+\lambda}{\sum_c [P(\theta_c \mid \psi_i)+\lambda]}$

G. Modified Training and Testing Process

The preprocessing for training and testing processes consists of (1) Detect condense *KP* features for training images, (2) Collect *KP* feature vectors and use online *spherical k-means algorithm* to find the representative descriptions. (3) Determine the number of neighbors N. The denotations are (1) k is the center numbers for k-means algorithm, (2) β is the scale factor of the neighboring *SR*s, (3) σ is the decay factor of Gaussian-distributed weighting kernel, and (4) χ is the threshold for determining similar scale adjacent *SR*s.

Two-class Object Model Training:

For each training image:

(*i*) Calculate *BoF* histogram for *KP* features using soft-weighting assignment method.

(*ii*) For each *SR* (A_{SR_i}), calculate h_{SR_i} using Eq. (13).

(*iii*) For every two *SR*s,

(*a*) If Eq. (11) is satisfied, then they are adjacent *SR*s.

(*b*) If $\chi^{-1} \leq R_{SR_i}/R_{SR_j} \leq \chi$, then they are similar scale, otherwise, different scale.

(*iv*) Generate $f_{BPLO_{i,j}}^{similar}$ for training data of similar scale, and generate $f_{BPLO_{i,j}}^{different}$ for training data of different scale.

(*v*) Define $l_{BPLO_{i,j}} = \begin{cases} 1 \; if \; c_f = c_\theta \\ 0 \; if \; c_f \neq c_\theta \end{cases}$.

Develop the object models $\theta_c^{similar}$ and $\theta_c^{different}$ by using random forest algorithm for model regression.

Testing:

For each testing image and every object class models $\theta_c^{similar}$ and $\theta_c^{different}$:

(*i*) Calculate *BoF* histogram for every dense *KP* features using soft-weighting assignment method.

(*ii*) For each *SR* (A_{SR_i}), calculate h_{SR_i} using Eq. (13).

(*iii*) For every two *SR*s,

(*a*) If Eq. (11) is satisfied, then they are adjacent *SR*s.

(*b*) If $\chi^{-1} \leq R_{SR_i}/R_{SR_j} \leq \chi$, then they are similar scale, else they are different scales.

(*iv*) Compute $P\left(\delta(A_{SR_i}, A_{SR_j}) \middle| A_{SR_i}, A_{SR_j}, \theta_c^{similar}\right)$ if both (*a*) and (*b*) are satisfied.

(*v*) Compute $P\left(\delta(A_{SR_i}, A_{SR_j}) \middle| A_{SR_i}, A_{SR_j}, \theta_c^{different}\right)$ if only (*a*) is satisfied.

(*vi*) Calculate $P(\boldsymbol{X} \mid \boldsymbol{A}, \theta_c)$ by summing all probabilities of pairwise local observations as

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

where $\Psi^{similar}$ and $\Psi^{different}$ are the two training data sets of pair *SR*s which are of similar and different scale respectively.

(*vii*) Categorize the image by using maximum likelihood estimation.

VI. Experimental Results

In the experiment, we test our system using Caltech101 [14] and Scene-15 [26] datasets. First, we show the confusion matrix of classification results for Scene15 dataset and every 20 categories of Caltech101. The major concern for implementing the random forests is the memory space. The memory space limitation makes it impossible to test the entire dataset at the same time. So, we split Caltech101 dataset into 5 partitions in the experiments. Our experimental results are satisfactory comparing to *BoF* method and the other conventional models. Second, we shows some examples to demonstrate the effectiveness of *NBPLO*.

A. Caltech-101 dataset

Caltech101 database contains 31~800 images per class of real-world photos and man-made photos as shown in Fig. 6. Our experiments show that our model 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.

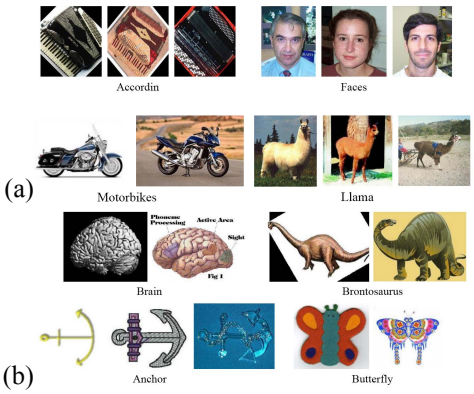


Fig. 6. (a) The real-world photos and (b) the artificial pictures.

Some correct classification examples are shown in Fig. 7. If more than half of local pairwise observations are correctly classified, then the image classification is correct. The class ambiguity happens because of similar-scale pairwise local observations. It is important to train the object model based on different-scale pairwise local observations. As shown in Fig. 7(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.

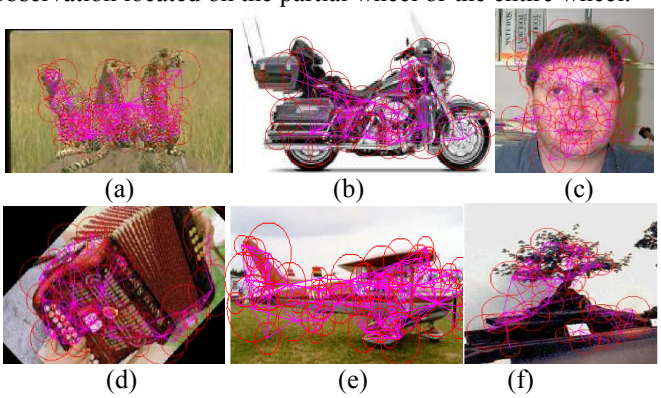


Fig. 7. Purple lines connecting two *SR*s shows that these pairwise local observations are assigned to correct class, whereas, blue lines connects the pairwise local observations of ambiguous category.

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 separate the low classification accuracy classes and the high classification accuracy classes for another testing. Each class contains more than 30 testing images for more accurate verification.

The confusion table of image classes with low classification accuracy is shown in Table 1. 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 of different information and different shapes are used to generate incoherence pairwise local observations.

The confusion table of image classes with high classification accuracy is shown in Table 2. The average accuracy of the well-classified images 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. Our method has two advantages: (1) random forest for likelihood prediction provides a good regression surface, and (2) the connectivity of pairwise local observations provides discriminative features.

B. Scene-15 dataset

Scene-15 database contains 15 different type of scene, with thirteen classes [26] (eight classes originate from [32]) and two other classes collected by [4]. Using the same setup as [4, 26], we randomly choose 100 images per class as the training data, and the others as the testing data, with the vocabulary size = 400. The accuracy rate of our method is 69.68%, which is better than standard *BoF* methods (50.55%), and the other results in [26] (65.2%). The results of our method is slightly lower than using only the first layer for matching pursuit [4] (72.2%). The confusion table of the image classification is shown in Table 3.

Table. 1. Confusion table of image classes with low classification accuracy.

buddha(55)	0.64	0	0.02	0.07	0	0	0.11	0.02	0.06	0.07	0	0.02	0
butterfly(61)	0	0.48	0	0.03	0.03	0.02	0.23	0.05	0.05	0.05	0	0.07	0
chair(32)	0.09	0	0.63	0	0.03	0.03	0.06	0	0	0.16	0	0	0
cougar_body(17)	0	0.12	0	0.24	0.12	0	0.12	0.06	0.24	0	0	0	0.12
crayfish(40)	0.07	0.15	0	0.03	0.45	0	0	0.13	0.03	0.1	0.03	0	0
dragonfly(38)	0.03	0	0.05	0.05	0.03	0.32	0.18	0.08	0.03	0.13	0.05	0	0.05
ewer(55)	0.09	0.05	0	0.04	0.05	0.02	0.44	0.02	0.04	0.25	0	0	0
hedgehog(24)	0	0.04	0	0.08	0.08	0	0	0.54	0.21	0	0.04	0	0
kangaroo(56)	0.05	0.04	0	0.25	0.02	0	0.09	0.09	0.43	0	0.02	0	0.02
lamp(31)	0.16	0.03	0.13	0	0	0	0	0.03	0.1	0.52	0	0	0.03
starfish(56)	0	0.02	0	0.05	0.05	0	0	0.2	0.04	0.07	0.48	0.02	0.07
stegosaurus(29)	0.07	0.07	0	0.14	0	0	0.03	0.1	0.03	0.03	0.03	0.45	0.03
umbrella(45)	0.07	0.02	0.11	0	0.07	0	0	0.02	0.02	0.2	0.04	0.02	0.42
	buddha	butterfly	chair	cougar_body	crayfish	dragonfly	ewer	hedgehog	kangaroo	lamp	starfish	stegosaurus	umbrella

Table. 2. Confusion table of image classes with high classification accuracy.

Face_easy(405)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Leopards(170)	0	0.98	0	0	0	0	0	0.01	0.01	0	0	0.01	0	0	0.01
Motobikes(768)	0	0.01	0.94	0	0	0	0.02	0	0	0	0	0	0	0	0.02
accordion(25)	0	0	0	0.92	0	0	0	0	0	0	0	0.04	0.04	0	0
airplanes(770)	0	0	0	0	0.83	0	0.16	0	0	0	0	0	0	0	0.01
camera(20)	0	0	0	0	0	0.65	0.05	0	0	0.05	0.1	0	0	0	0.15
car_side(93)	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
dalmatian(37)	0	0	0	0	0	0	0	0.97	0	0	0	0	0	0	0.03
dollar_bill(22)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
headphone(12)	0	0	0	0	0	0	0	0	0	0.92	0	0	0	0.08	0
ketch(84)	0.01	0.04	0	0	0.02	0	0.02	0	0.05	0.01	0.5	0	0	0.31	0.04
minaret(46)	0	0.17	0	0	0	0	0	0	0	0	0	0.83	0	0	0
pagoda(17)	0	0	0	0.12	0	0	0	0	0	0	0	0	0.88	0	0
schooner(33)	0	0	0	0	0.03	0.03	0	0	0.06	0	0.27	0	0	0.55	0.06
wheelchair(29)	0	0	0	0	0	0.03	0	0	0	0	0	0	0	0	0.97
	Face_easy	Leopards	Motobikes	accordion	airplanes	camera	car_side	dalmatian	dollar-bill	headphone	ketch	minaret	pagoda	schooner	Wheel-chair

Table 3. Confusion table of image classification of Scene-15 dataset.

CALsuburb(141)	0.96	0	0	0	0	0	0	0	0	0.01	0	0	0.01	0.01	0.01
MITcoast(260)	0.01	0.77	0	0.08	0	0.01	0.12	0	0	0	0	0	0	0	0
MITforest(228)	0	0	0.82	0	0	0.09	0.06	0.01	0	0	0	0	0	0	0.01
MIThighway(160)	0.01	0.04	0	0.86	0.01	0	0.03	0.01	0	0	0	0.03	0	0	0.01
MITinside-city(208)	0.02	0	0	0.01	0.67	0	0	0.06	0.01	0.02	0.01	0.02	0.06	0.01	0.1
MITmountain(274)	0	0.02	0.02	0.02	0	0.81	0.08	0.04	0	0	0	0	0	0	0
MITopencountry(310)	0.01	0.1	0.06	0.08	0	0.06	0.67	0.01	0	0	0	0	0	0	0
MITstreet(192)	0.02	0	0	0.02	0.07	0	0	0.72	0.03	0	0	0.03	0.02	0.03	0.06
MITtallbuilding(256)	0.01	0	0.02	0	0.04	0.02	0	0.04	0.75	0	0.02	0.04	0.01	0.02	0.05
PARoffice(115)	0.01	0	0	0	0	0	0	0	0	0.89	0.02	0	0.04	0.04	0
Bedroom(116)	0.02	0	0	0.01	0.02	0.02	0	0.01	0	0.06	0.47	0.03	0.14	0.2	0.03
industrial(211)	0.06	0	0	0.07	0.07	0	0.01	0.1	0.06	0.03	0.02	0.26	0.06	0.12	0.14
kitchen(110)	0.02	0	0	0	0.03	0	0	0.01	0	0.1	0.04	0	0.67	0.09	0.05
livingroom(189)	0.02	0	0	0	0.02	0	0	0.01	0	0.07	0.1	0.02	0.13	0.52	0.12

store(215)	0.01	0	0	0.02	0.05	0.05	0	0.04	0.01	0.02	0.01	0.03	0.04	0.08	0.63
	CAL suburb	MIT coast	MIT forest	MIT highway	MIT inside-city	MIT mountain	MIT open-country	MIT street	MIT tall-building	PAR office	Bed-room	industrial	kitchen	Living-room	store

Then, we show some correctly classified images most of which include the dominant objects in the scene. In Fig. 8(a), the *SR*s contain partial appearance of tall buildings. Figs. 8(b) and 8(c) show that most of the pairwise observations are recognized as offices or bedroom. In Fig. 8(d), the dominant local observations located on most of the products on the shelves belong to the class "store", while the ambiguous local observations concentrate at the locations of the lamp and the windows. In Fig. 8(e), local observations of the ocean connect to local observations of the sands are classified as "beach", and observations of the trees, sky, and clouds are ambiguous. In Fig. 8(f), pairwise local observations of trees are classified as "forests", and the ingredients of the house are ambiguous. In Fig. 8(e), apparently ambiguous pairwise local observations may be classified as "mountain", however the majority pairwise local observations are categorized to "beach". In Fig. 8(*h*), the ambiguous pairwise local observations are distributed through the grounds, while the local observations that capture partial appearance of trees are correctly classified.

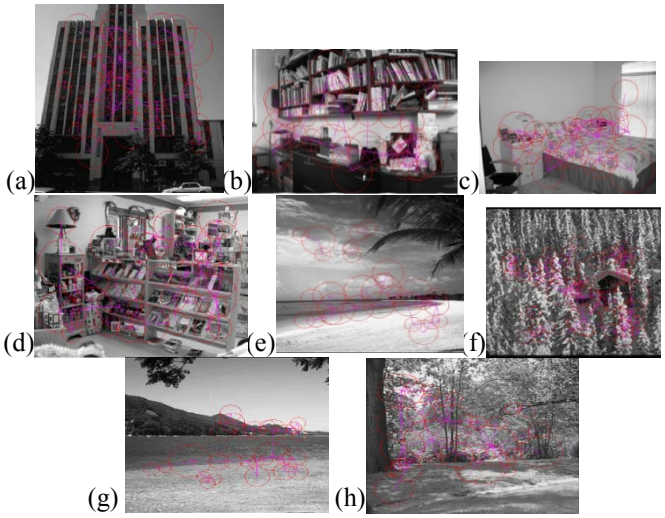


Fig. 8. The results of correct classified images.

The above examples show that our model is effective. Our naïve-Bayes assumption is that a large proportion of components belong to the designate class. This method may fail to classify the images containing objects that belong to the other classes, such as the image contains house in the forest or a beach surrounded by mountains. Next, we demonstrate some miss-classified images shown in Fig. 9. The image shown in Fig. 9(a) is miss-classified as "kitchen". Most of the decorations do not clearly belong to bedroom, and only one-fourth of the image contains bed and sofa. The image shown in Fig. 9(b) belongs to the class "industry". Most of the *SR*s located at the clouds. Obviously, most of the *SRs* do not contain any information regarding to the "industry". Figs. 9(c) shows that the image is classified as "office". It seems reasonable, because most of the information are hairs and windows, which consist insufficient information for the system to recognize it as "kitchen". Combining *KP* feature with *SR*s may fail if the *SR*s do not capture sufficient information of the images such as Fig. 9(d). The total area of *SR*s does not exceed one-third of the total image area.

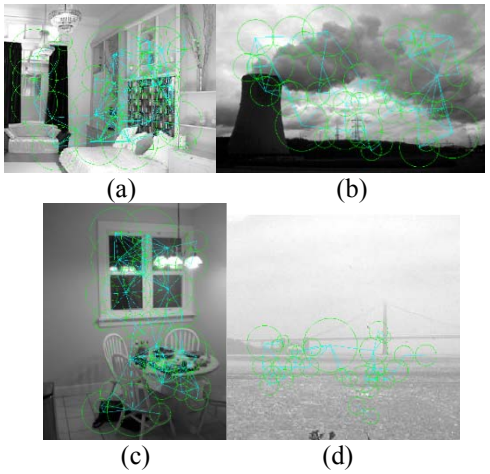


Fig. 9. The results of miss-classified images.

VII. Conclusions

We have proposed an image classification method based on local pairwise observations without any priori information. Different from the pyramid matching pursuit, our method also outperforms the conventional *BoF* method. However, there are still more room for improvement and some problems to be solved. We may improve our naïve-Bayes assumption, which is too "naive" for general image recognition. Second, the memory space problem prevent us from doing experiments on massive database.

References

[1], D. G. Lowe, "Distinctive image features from scale-invariant *KP*s." *Int. Journal of Computer Vision Vol. 60(2), pp.91-110, 2004.*

[2] B. Herbert, T. Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features." *ECCV 2006. 404-417.*

[3] C. Harris and M. Stephens, "A combined corner and edge detector." *Alvey Vision Conference. Vol. 15. 1988.*

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching For Recognizing Natural Scene Categories." *IEEE CVPR, Vol. 2. 2006.*

[5]J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification." *IEEE CVPR, 2009.*

[6] J. Wang, J. Yang, K.Yu, F. Lv, T. Huang, and Y. Gong, "Locality-Constrained Linear Coding For Image Classification." *IEEE CVPR 2010.*

[7] Z. Jiang, Z. Lin, and L. S. Davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition." *IEEE Trans. on PAMI, 35(11), 2013.*

[8] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition." *IEEE Trans. on PAMI 28(12), 2006.*
[9] T. Jabid, M H Kabir, and O. Chae, "Local directional pattern (LDP) for face recognition." *ICCE, 2010.*
[10] Y. Jia, C. Huang, and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features." *IEEE CVPR, 2012.*
[11] M. C. Burl, M. Weber, and P. Perona.,"A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry." *ECCV, pp. 628-641, 1998.*
[12] R. Fergus, Pietro Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning." *2003 IEEE CVPR, Vol. 2.*
[13] T. Kadir and M. Brady, "Saliency, scale and image description." *Int. J. of Computer Vision, pp.83-105, vol. 45, no.2, 2001.*
[14] 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 106.1 pp.59-70, 2007.*
[15] K. Mikolajczyk, "Detection of local features invariant to affine transformations." *Ph.D. thesis, Institut National Polytechnique de Grenoble, France,* 2002.
[16] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun Database: Large-Scale Scene Recognition from Abbey to Zoo"*, IEEE CVPR, 2010.*
[17] Juan, Luo, and Oubong Gwun, "A comparison of SIFT, PCA-SIFT and SURF." *Int. Journal of Image Processing (IJIP) 3.4 (2009): 143-152.*
[18] Liu, Ce, Jenny Yuen, A. Torralba, Josef Sivic, and W. T. Freeman, "Sift flow: Dense Correspondence across Different Scenes." *ECCV 2008.*
[19] Otero, Ives Rey, and Mauricio Delbracio. "Anatomy of the SIFT Method." (2013).
[20] A. Vedaldi and B.Fulkerson, "VLFeat: An open and portable library of computer vision algorithms." *Int. Conf. on Multimedia. ACM, 2010.*
[21] C-F Tsai,"Bag-of-words representation in image annotation: A review." *ISRN Artificial Intelligence 2012.*
[22] J. Yang, Y-G Jiang, A. G. Hauptmann, and C-W Ngo, "Evaluating bag-of-visual-words representations in scene classification." *Int. Workshop on Multimedia Information Retrieval. ACM, 2007.*
[23] S. Zhong "Efficient online spherical *k*-means clustering." *IEEE IJCNN, Vol. 5, 2005.*
[24], Y-G. Jiang, C-W. Ngo, and J. Yang, "Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval." *The 6th ACM Int. Conf. on Image and Video Retrieval, 2007.*
[25] F. Pedro, D. McAllester, and D. Ramanan. "A discriminatively trained, multiscale, deformable part model." *IEEE CVPR 2008.*
[26] F. F. Li, and P. Perona. "A Bayesian Hierarchical Model for Learning Natural Scene Categories." *IEEE CVPR 2005.*
[27] L. Breiman, "Random forests." *Machine learning 45.1 (2001): 5-32.*
[28] J.R.R. Uijlings, A.W.M. Smeulders, and R.J.H. Scha. "What is the Spatial Extent of an Object ?." *IEEE CVPR, 2009.*
[29] J. Wang, , Y. Li, Y. Zhang, C.Wang, H. Xie, G. Chen, and X. Gao. "Bag-of-Features Based Medical Image Retrieval Via Multiple Assignment and Visual Words Weighting." *IEEE Trans. on Medical Imaging 30, no. 11, 2011.*
[30] J. Lafferty, A. McCallum, and F. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." *ICML 2001*, pages 282-289.
[31] A. Bosch, A. Zisserman, and X. Muoz,"Image Classification using Random Forests and Ferns," *ICCV 2007.*
[32] A. Oliva and A. Torralba,"Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope." *Int. Journal of Computer Vision 42(3), 2001.*
[33] K. Kesorn and S. Poslad, "An Enhanced Bag-Of-Visual Word Vector Space Model to Represent Visual Content in Athletics Images." *IEEE Trans. on Multimedia, 14(1), 2012.*