# Object Classification based on Weakly Supervised E2LSH and Saliency map Weighting

**Yongwei Zhao[1], Bicheng Li[1], Xin Liu[1], Shengcai Ke[1]**
1 China National Digital Switching System Engineering and Technological R&D Center
Zhengzhou, Henan 450002 - P. R. China
[e-mail: zhaoyongwei369@163.com, lbclm@163.com, Liuxin1015@163.com, keshengcai0705@163.com]
*Corresponding author: Yuansong Li

*Received July 10, 2015; revised September 30, 2015; accepted October 20, 2015; published January 31, 2016*

## Abstract

The most popular approach in object classification is based on the bag of visual-words model, which has several fundamental problems that restricting the performance of this method, such as low time efficiency, the synonym and polysemy of visual words, and the lack of spatial information between visual words. In view of this, an object classification based on weakly supervised E2LSH and saliency map weighting is proposed. Firstly, E2LSH (Exact Euclidean Locality Sensitive Hashing) is employed to generate a group of weakly randomized visual dictionary by clustering SIFT features of the training dataset, and the selecting process of hash functions is effectively supervised inspired by the random forest ideas to reduce the randomcity of E2LSH. Secondly, graph-based visual saliency (GBVS) algorithm is applied to detect the saliency map of different images and weight the visual words according to the saliency prior. Finally, saliency map weighted visual language model is carried out to accomplish object classification. Experimental results datasets of Pascal 2007 and Caltech-256 indicate that the distinguishability of objects is effectively improved and our method is superior to the state-of-the-art object classification methods.

***Keywords:*** Object Classification; Bag of Visual Words; E2LSH; Graph-based Visual Saliency; Visual Language Model Method;

This research was supported by the National Natural Science Foundation of China [Grant No. 60872142 and Grant No. 61301232 ].

**http://dx.doi.org/10.3837/tiis.2016.01.021** **ISSN : 1976-7277**

## 1. Introduction

With the explosive growth of the images, object classification has been an important issue in computer vision owing to its great potential in both research problems and industry applications. The appearance of bag of visual words (BoVW) model [1-5] has taken the first step toward transition from low level visual features to high level semantic concepts; it is currently the mainstream method in the image classification field. The BoVW model first generate a visual dictionary by clustering local features [6] of the training images into different clusters with the centroids as visual words, where the clustering algorithms typically is k-means [1]. Then, the SIFT features extracted from a test image are mapped to visual dictionary for constructing a visual vocabulary histogram which are used for image classification by some machine learning methods.

However, there are some drawbacks in the traditional BoVW model that limiting its performance. Firstly, it is low time efficiency in visual vocabulary construction part. Most existing algorithms (e.g. K-Means) clustering as it involves several distance calculations of each data point from all the centers in each iteration. Thus, as the visual data scales up, the time efficiency decreases rapidly. Secondly, the initial cluster centers are randomly generated, thus the clustering results are unstable and susceptible to noise data points, which lead to the ambiguity of visual words [7], which lies in two areas: synonymy and polysemy. Moreover, the traditional BoVW model suffers from another main problem is the loss of spatial information when representing the images as histograms of quantized features and suffers limited accuracy[8].

In order to improve the efficiency of dictionary construction, reference [9] exploited a nested structure of Voronoi cells for K-Means, known as Hierarchical k-means (HKM). In [10] it was shown that this reduced time complexity could also be achieved by a KD-forest approximation. Zhang [11] proposed an E2LSH-MKL based dictionary construction method. This method utilizes nonlinear combination of multiple different kernels in order to make full use of information generated from the nonlinear interaction of different kernels. The method in [11] can demonstrate superior time and space efficiency over K-Means and HKM or AKM, in both theory and practice. But, the hash functions in [11] are randomly generated without the prior information of the training data. It will undoubtedly reduce the representation and distinction of the generated dictionaries.

In order to overcome the influence of this negative factor brought by synonymy and ambiguity of visual words, many researchers have made lots of explorations and attempts. One possible way to disambiguate the visual words ambiguity is to combine visual words into a larger unit, a so-called visual sentence. [12], [13] proposed a solution to tackle this issue using the “visual phrase”. The major weakness of the visual phrase approach is that it only considers the co-occurrence information among visual words but neglects spatial information amongst the visual words. Philbin et al. [14] presented a kind of BOVW model based on soft-assignment to build the visual vocabulary histogram. In which, a SIFT feature point is assigned to several weighted nearest visual words rather than hard-assignment [9] to a single word as in previous work. Gemert et al. [7]established a visual word uncertainty model, where some kernel functions were carried out to complete soft-mapping between local features and visual words, this model can efficiently decrease the quantization error, and the results further verify the effectiveness of soft assignment method in solving the synonymy and ambiguity problem of visual words. Wang et al [15] proposed a QP (quadratic programming) assignment method by using the linear construction weights of the neighboring visual words as the contribution functions. The weights can be formulated and solved by the quadratic programming techniques.

Moreover, many efforts have been made to capture the spatial information of visual words. Spatial Pyramid Matching (SPM) [16] has been proven a simple but effective extension to bag-of-visual-words image representation for spatial layout information compensation by partitioning images into coarse-to-fine sub-blocks and concatenating the histograms extracted from all blocks. Sharma et al. [17] proposed spatial-bag-of-features to extend SPM by introducing two different ways for image partition. Xie et al [18] proposed a novel pooling algorithm named Geometric Phrase Pooling (GPP) to capture the spatial contexts. However, target objects may appear at any location in the image with various backgrounds. Therefore, the fixed spatial matching in SPM fails to match similar objects located different locations and backgrounds. Besides, the methods based on SPM will suffer from high computational cost in real-time application since they need enumerating a huge number of possible line angles and center locations. In addition, the visual language model (VLM)[19][20]has become a popular approach to tackle the spatial information loss of visual words due to its good empirical success. A visual language model utilizes a probability distribution that captures the statistical properties of visual words to express image semantic content. However, not all regions on a natural image are really useful for object classification; the visual words belonging to the background region will bring in some noises. Nakamoto et al [21] proposed to use a binary mask, derived from the saliency map, to remove the visual words that appear in the irrelevant part of the image. The methods in [21] tend to discard or reduce the influence of the background in the object classification process, But, it can not to weight and analyze the relevance of visual words in background and foreground.

To overcome the above mentioned problems, we propose an object classification method based on weakly supervised E2LSH and saliency map weighting. In our method weakly supervised E2LSH is employed to generate a group of weakly randomized visual dictionary by clustering SIFT features of the training dataset. E2LSH [22] is a scheme of Locality Sensitive Hashing [23] realized in Euclidean space. The basic idea of E2LSH is that several locality sensitive hashing functions is used to map high-dimensional data points into low-dimensional space ensuring the closer points in initial space still close to each other. Here, inspired by random forest algorithm, a new weakly supervised scheme is proposed to choose out the best hash functions, which can effectively reduce the randomness of traditional E2LSH algorithm and enhance the distinction and extensibility of visual dictionaries. Then, Graph-Based Visual Saliency (GBVS) algorithm [24] is applied to detect the saliency map of images and weight the visual words according to the saliency prior. Finally, saliency map weighted visual language model is carried out to accomplish object classification.

The remainder of this paper is organized as follows. In Section 2 we briefly introduce the E2LSH algorithm. In Section 3, we introduce object classification based on weakly supervised E2LSH clustering and saliency map weighted visual language model in detail. We present the experimental results in Section 4. Finally we conclude this paper in Section 5.

## 2. Exact Euclidean Locality Sensitive Hashing Overview

Locality sensitive hashing is first widely used to solve the large-scale and rapid image similarity search problem using several hash functions to ensure the probability of collision is much higher for objects which are close to each other than for those which are far apart after a “projection” operation. The hash functions applied in E2LSH are based on p-stable distributions [22]. Here, all the employed LSH functions are 2-stable and defined as,

$$h_{\alpha,\beta}(\boldsymbol{v}) = \left\lfloor \frac{\boldsymbol{\alpha}\cdot\boldsymbol{v}+\beta}{\varpi} \right\rfloor \tag{1}$$

Where, $\lfloor\cdot\rfloor$ is the flooring function, $\boldsymbol{a}$ is a $d$ -dimensional vector with components that are selected randomly from $p$ -stable distribution and $\beta$ is a random variable uniformly distributed in $[0,\varpi]$. $\varpi$ denotes a constant. Each hash function $h_{a,b}(\boldsymbol{v}): \mathrm{R}^d \to \mathrm{Z}$ : maps a $d$ dimensional vector $\boldsymbol{v}$ onto the integer set. However, a single hash function is not discriminative enough, so E2LSH usually combines several LSH functions by defining a family $\mathrm{G} = \{g: S \to U^k\}$ , where $g(\boldsymbol{v}) = (h_1(\boldsymbol{v}), \cdots, h_k(\boldsymbol{v}))$ . For each point $\boldsymbol{v} \in \mathrm{R}^d$ , $\boldsymbol{a} = (a_1, a_2, \cdots a_k)$ can be obtained after mapping through $g(\boldsymbol{v}) \in \mathrm{G}$. Then, the primary and the secondary hash function $h_1$ $h_2$ are applied to hash the vector $\boldsymbol{a}$ and establish hash tables for saving data points. $h_1$ $h_2$ are defined as,

$$h_1(\boldsymbol{a}) = ((\sum_{i=1}^{k} r_i' a_i) \bmod m) \bmod s \tag{2}$$

$$h_2(\boldsymbol{a}) = (\sum_{i=1}^{k} r_i'' a_i) \bmod m \tag{3}$$

Where $r_i'$ and $r_i''$ are random integers, $s$ denotes the size of the hash tables, and $m$ is a prime number, which equals $2^{32}-5$. The data points with the same $h_1$ and $h_2$ value will be hashed into the same bucket in hash table, thus realizing these data points' partition.

Based on this, E2LSH can be employed to perform visual dictionary' construction, if each bucket center of the hash table is viewed as a visual word, the hash tables can be seen as visual dictionaries., we propose a weakly supervised scheme to weaken the randomicity of E2LSH. The basic idea is to choose out the best hash functions using the training data, which can effectively enhance the distinction and representation of hash functions. Moreover, we select $L$ dependent $g_1, \cdots, g_L$ from $\mathrm{G}$ to further reduce the randomness of generated visual dictionaries.

## 3. Object Classification based on weakly supervised E2LSH and Saliency map weighting

For training image dataset $\mathrm{I} = \{\mathrm{I}_1, \mathrm{I}_2, \ldots \mathrm{I}_k\}$, the entire process of object classification based on weakly supervised E2LSH and saliency map weighting as illustrated in **Fig. 1**. The following we will introduce the weakly supervised E2LSH clustering algorithm and saliency map weighted visual language model in detail.

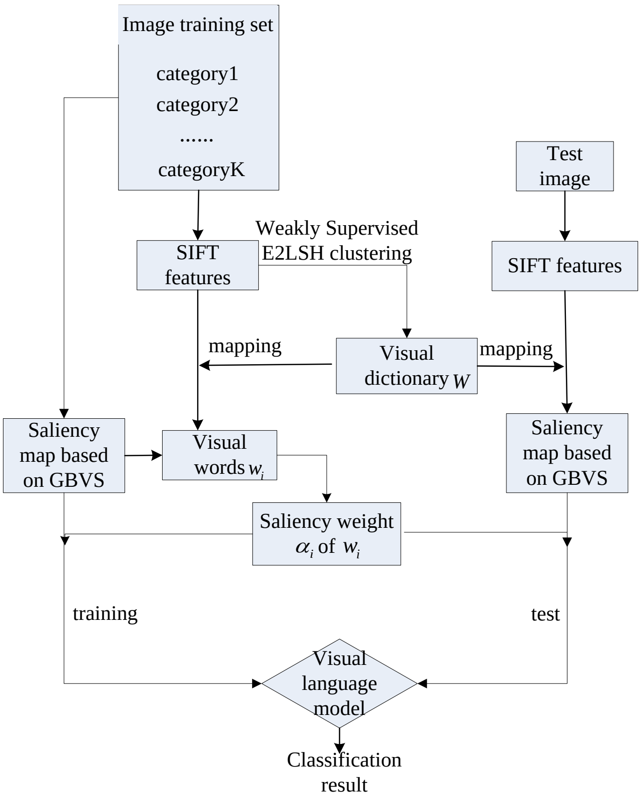


**Fig. 1.** The flow chart of object classification based on weakly supervised E2LSH and saliency map weighting

### 3.1 Weakly supervised E2LSH clustering

Random Forests has been proposed by Breiman [25] as an enhancement of tree bagging. This algorithm can estimate the relevance between trees, the smaller correlation with the stronger distinction. So, inspired by this idea, a new weakly supervised scheme is proposed to choose out the best hash function, which can effectively reduce the randomness of traditional E2LSH algorithm. Here, we suppose $j$ hash functions $h_1, h_2, ...h_j, 1 \le j < k$ have been selected, the weakly supervised selection process of $(j+1)$-$th$ hash function can be depicted as follows,

The weakly supervised selection process:

Step1: Use the function $g_i$ including $j$ hash functions $h_1, h_2, ...h_j$ to do reduction mapping for visual SIFT vector $R = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_j, \cdots, \boldsymbol{r}_{N_s}\}$ and get a $j$ -dimensional vector $g_i(\boldsymbol{r}), 1 \le i \le L$, where $L$ is the number of hash tables , $\boldsymbol{r}_{N_s}$ is the total number of SIFT features, $g_i(\boldsymbol{r})$ is the dimensionality reduction vector of the SIFT feature $\boldsymbol{r}$ ;

Step2: Calculate the primary and secondary hash key $h_1(g_i(\boldsymbol{r}))$, $h_2(g_i(\boldsymbol{r}))$ according to Eqs (2) and (3) respectively. The SIFT points with the same $h_1(g_i(\boldsymbol{r}))$ and $h_2(g_i(\boldsymbol{r}))$ will be hashed into the same bucket, which can be seen as a visual word, the visual dictionary $W(w_1, w_2, ...w_{N_j})$ is generated, $N_j$ is the dictionary size;

Step3: Calculate Shannon entropy of each visual word $w_i$ that in dictionary $W$ according to $H_C(w_i) = -\sum_{l \in k} \frac{n_l}{n} \log_{w_{ij}} \frac{n_l}{n}$ , where, $n$ is the number of SIFT features in the $i$-$th$ word $w_i$ , $n_l$ denotes the number of SIFT features that belonging to $l$ object category in $i$-$th$ word;

Step4: Choose a new $h$ function as the candidate $(j+1)$ hash function $\hat{h}$ and calculate the split entropy $H_S(\hat{h}, w_i) = -\sum_{j=1}^{w_{ij}} \frac{n_j}{n} \log_{w_{ij}} \frac{n_j}{n}$ of $i$-$th$ bucket, here, we suppose the initial $i$-$th$ bucket will be split into $w_{ij}$ buckets by the candidate hash function $\hat{h}$, and $n_j$ is the number of SIFT features in $w_{ij}$-$th$ word. $H_S(\hat{h}, w_i)$ will reach the maximum when the $b_{ij}$ partitions have equal size;

Step5: Based on the step3 and step4, the impurity of word $w_i$ can be calculated by the mutual information according to $I_i(\hat{h}) = H_C(w_i) - \sum_{j=1}^{w_{ij}} \frac{n_j}{n_i} H_C(w_{ij})$ . Here, we calculate a score as $S_{j+1}(\hat{h}) = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{2I_i(\hat{h})}{H_C(w_i) + H_S(\hat{h}, w_i)}$ for each candidate $\hat{h}$ to choose the best $(j+1)$-$th$ hash function;

Step6: Repeat step4-step5, we will get several candidate hash functions $\hat{h}$, and we can select the best one according to $h^* = \arg\max_{\hat{h}} S_{j+1}(\hat{h})$ .

Repeat the above mentioned process; we can get $k$ best hash functions with stronger representation and distinction, which can efficiently weaken the randomicity of E2LSH. In order to further reduce the randomicity, we choose $L$ independent functions $g_1, \cdots, g_L$ to construct a group of weakly randomized dictionaries. The flow can be described as **Fig. 2**,

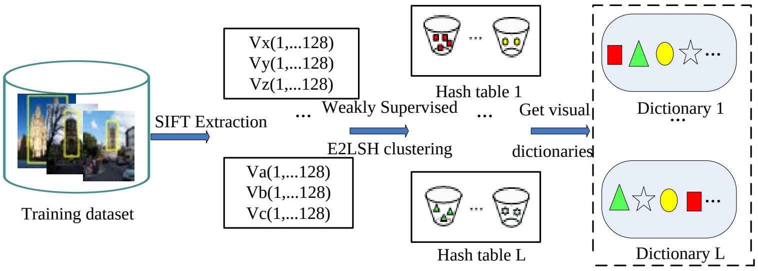


**Fig. 2.** The construction process of weakly randomized visual dictionaries

Step1: Features extraction. Extract the SIFT features for all images according to the method of reference [6] to obtain the visual feature vector $R = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_j, \cdots, \boldsymbol{r}_{N_s}\}$, where $\boldsymbol{r}_j$ is a

128-dimensional SIFT descriptor, $N_s$ denotes the total number of SIFT descriptors . For a SIFT point $\boldsymbol{r} \in R$ , we employ the function $g_i$ to do reduction mapping and get a $k$ -dimensional vector $g_i(\boldsymbol{r}), 1 \le i \le L$ ;

Step2: Weakly supervised E2LSH hash clustering. Calculate the primary and secondary hash key $h_1(g_i(\boldsymbol{r})), h_2(g_i(\boldsymbol{r}))$ according to Eqs (2) and (3) respectively. The SIFT points with the same $h_1$ , $h_2$ will be hashed into the same bucket to generate hash tables $T_i = \left\{b_1^{(i)}, b_2^{(i)}, \cdots, b_k^{(i)}, \cdots, b_{N_i-1}^{(i)}, b_{N_i}^{(i)}\right\}, 1 \le i \le L$ , where $b_k^{(i)}$ denotes the number $k$-$th$ bucket in $T_i$ and correspond to a cluster, which can be seen as a visual word, $N_i$ is the hash table size.

Step3: Visual words elimination. The visual words with little discrimination can be removed according to the weight calculated by Eqs (4) from small to large, here, we save $M$ words for each visual dictionary, that is $W_i = \left\{w_1^{(i)}, w_2^{(i)}, \cdots, w_k^{(i)}, \cdots, w_{M-1}^{(i)}, w_M^{(i)}\right\}$, $i = 1, \cdots, L$.

Obviously, the whole process is scalable. So, if a new image is added in, we only need to detect its' SIFT points and hash these points by weakly supervised E2LSH to realize the dynamic expansion of randomized visual dictionaries.

The benefits of the weakly supervised E2LSH for overcoming the problem of synonymy and polysemy of visual words are firstly illuminated by **Fig. 3**. In **Fig. 3**, points $w_1$—$w_5$ represent cluster centers and points 1—5 are SIFT features, the features 3,4,5 are close in image space and can be assumed they represent the same image content area. On the contrary, features 1, 3 are far from each other, and can be assumed they represent the different image content area. In hard assignment [8], features 3, 4 and 5 will be assigned to different visual words despite being close in feature space, that is the synonymy phenomenon of visual words $w_1, w_3, w_2$ . The features 1 and 3 are assigned to word $w_1$ equally, that is the polysemy phenomenon of visual word $w_1$ . But using weakly supervised E2LSH mapping, features 3, 4 and 5 may be hashed to one bucket with a high probability, features 1 and 3 may hardly be hashed to a same bucket, meanwhile, because of the weak randomcity of E2LSH, when we use the multi-hash tables to generate dictionaries, it can also make a same feature be mapped to some multi-words like soft assignment method [14]. So if we employ weakly supervised E2LSH to perform feature points' clustering and mapping, it not only can overcome the low time efficiency of K-Means but also solves the problem of synonymy and polysemy of visual words well.

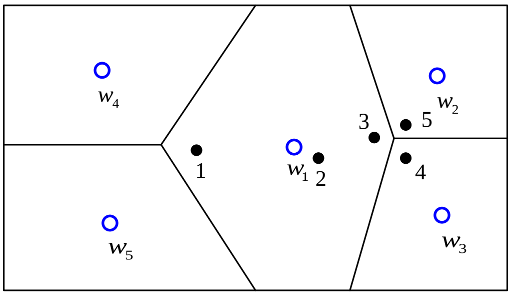
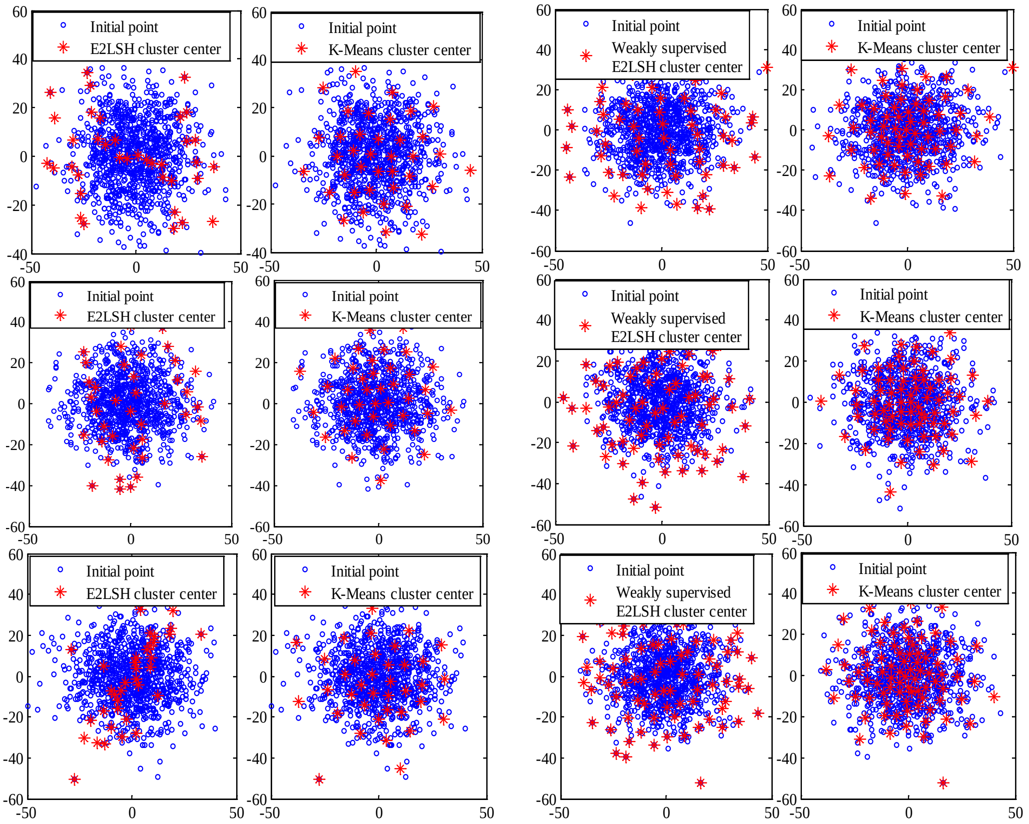


**Fig. 3.** Benefits of the weakly supervised E2LSH

In order to prove the effectiveness of weakly supervised E2LSH clustering for dataset, we compare weakly supervised E2LSH, E2LSH and K-Means clustering algorithms on the same randomly generated data points in Matlab 2012 circumstance. The cluster results are as shown in **Fig. 4**(a) and **Fig. 4**(b). The blue circles represent the original data points and the red star points are cluster centers produced by weakly supervised E2LSH, E2LSH and K-Means respectively. From **Fig. 4**(a) we can see that there is stronger randomness in cluster results of E2LSH, which is not superior to the cluster results of K-Means. From **Fig. 4**(b) we can see that the weakly supervised scheme can effectively reduce the randomicity of E2LSH and enhance the robustness of its clustering results. Moreover, from **Fig. 4**(a) and **Fig. 4**(b), we can see that K-Means has more cluster centers in dense areas but less cluster centers in sparse areas. However, the cluster centers' distribution of weakly supervised E2LSH is more uniform and more stable, which is conducive to overcome the problem of ambiguity of visual words.



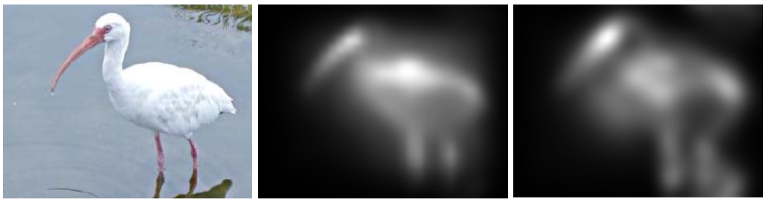
(a) (b)

**Fig. 4.** Comparison of clustering results of E2LSH, weakly supervised E2LSH and K-Means

## 3.2 Saliency map weighted visual language model

We are more interested in the objects appear on different backgrounds in object classification work. In view of this, we propose a saliency map weighted visual language model to classify object images. Itti et al. [26] proposed an saliency map detection model purely bottom-up in which the input scene is decomposed into a set of topographic feature maps, that are intensity maps, color maps and orientation maps. Harel et al. [24] proposed a bottom-up saliency map which uses the Markov chains over various feature maps and treat the equilibrium distribution over map locations as activation and saliency values, which is prior to the approch proposed by

Itti in [26]. **Fig. 5** gives the different saliency map results derived from the same image by the methods in [24][26]. From **Fig. 5** it can be seen that the GBVS algorithm can obtain clearer object foreground than the method proposed by Itti. Therefore, we adapt the GBVS algorithm to detect saliency map of images and weight the visual words according to the saliency prior.



GBVS map Itti map

(a) original image (b) saliency maps

**Fig.5.** (a) Original image; (b) the saliency maps derived from the image shown in (a) using the approachs proposed by Harel [24] and Itti[26].

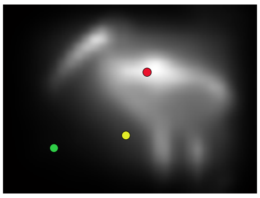


**Fig. 6.** SIFT points with different significance according to different positions

The three SIFT points highlighted in **Fig. 6** illustrate three different situations of information uncertainty: the red point highly belongs to the foreground, then assigns to it a high membership degree; the green point lowly belongs to the foreground, and then assigns to it a low membership degree. The yellow point is located in a transition region. Here, we weight the visual words $w_k^{(i)}$ as follows,

$$\alpha_k^{(i)} = \frac{\underset{\boldsymbol{r}_j\in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 1\right) + 1/2\,\underset{\boldsymbol{r}_j\in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 0\right)}{\underset{\boldsymbol{r}_j\in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j\right)} \qquad （4）$$

$$\beta_k^{(i)} = 1 - \alpha_k^{(i)} \qquad （5）$$

Where, $w_k^{(i)}$ denotes the $k$-$th$ visual word of dictionary $W_i, 1 \le i \le L; 1 \le k \le M$ , $\alpha_k^{(i)}$ , $\beta_k^{(i)}$ are the weights for foreground saliency and background saliency respectively, represented in the interval [0,1]. $\boldsymbol{r}_j \in w_k^{(i)}$ represents the number of SIFT points that hashed to the bucket of word $w_k^{(i)}$ , $y_j$ is the identify value of SIFT point $\boldsymbol{r}_j$ . $y_j = 1$ denotes $\boldsymbol{r}_j$ belonging to the foreground area as the red point in Figure 6, $\underset{\boldsymbol{r}_j\in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 1\right)$ denotes the number of SIFT points that hashed to $w_k^{(i)}$ and belong to the foreground area; $y_j = 0$ denotes the SIFT point locate in a transition region, $\underset{\boldsymbol{r}_j\in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 0\right)$ is the number of SIFT points that hashed to $w_k^{(i)}$ and belong to transition region as the yellow point in Figure 6; $y_j = \text{-}1$ represents the $\boldsymbol{r}_j$ belonging to background area like the green point in Figure 6. To avoid the zero probability, a simple smoothing method is proposed as follows,

$$\alpha_k^{(i)} = \frac{\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 1\right) + 1/2\,\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 0\right) + 1}{\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j\right) + 1} \tag{6}$$

Then, we use one gram visual language model to express image content as Equation (7), The simplest way to estimate the image language model is to treat the image as a sample from the underlying multinomial word distribution and use the maximum likelihood estimator (MLE) as Equation (8)，

$$p(w_k^{(i)} \mid w_0^{(i)} w_1^{(i)} \ldots w_M^{(i)}) = p(w_k^{(i)}) \tag{7}$$

$$p(w_k^{(i)} \mid C_t) = \frac{\alpha_k^{(i)} \cdot F_N(w_k^{(i)} \mid C_t)}{\sum_{w_k^{(i)} \in W_i} F_N(w_k^{(i)} \mid C_t)} \tag{8}$$

Where, $C_t$ is the $t$-$th$ object category, $F_N(w_k^{(i)} \mid C_t)$ denotes the count of the visual word $w_k^{(i)}$ in $C_t$ ,however, Equation (8) could generate a zero probability if a visual word $w_k^{(i)}$ never occurs in the document image. To avoid the incorrectness, the Jelinek-Mercer smoothing method [27] is motivated to linearly interpolate the maximum likelihood estimator and the collection one as follows,

$$p_\lambda(w_k^{(i)} \mid C_t) = (1-\lambda)\,p(w_k^{(i)} \mid C_t) + \lambda\, p(w_k^{(i)} \mid C) \tag{9}$$

Where, $p(w_k^{(i)} \mid C)$ is the visual word $w_k^{(i)}$ distribution in training image set $C$ , $\lambda \in [0,1]$ is a trade-off parameter to balance the contribution of MLE and collection model. As long as $\lambda > 0$ and the collection estimation $p(w_k^{(i)} \mid C) > 0$ , it can be seen that $p_\lambda(w_k^{(i)} \mid C_t) > 0$ no matter $w_k^{(i)}$ is seen or not in $C_t$ . Here, we utilize Eqs (10) to classify a testing image $I$ .

$$C^* = \arg\max_{C_t} \frac{1}{L} \sum_1^L \prod_{w_k^{(i)} \in I} p\left(w_k^{(i)} \mid C_t\right) p\left(C_t\right), i = 1,, 2, \ldots, L; t = 1, 2, \ldots, N_t \tag{10}$$

# 4. Experiments

## 4.1 Experimental dataset setup and evaluation

In this experiment, we use the standard test image collection Caltech-256 [28] and Pascal Voc 2007 dataset [29] to evaluate the performance of our method. Firstly, we do some experiments on the whole Caltech-256 dataset to evaluate the effectiveness of algorithms proposed in this article. 50 images are choose in each category to construct training image set for generating visual dictionary and the remaining are as testing set. The visual dictionary size is 20K. **Fig. 7** shows the sample images of each category. To obtain reliable experimental results, all image classification experiments are run 10 times and then averaged to produce the final average precision. The hardware configuration for experiment is a desktop with Core 3.1G×4 CPU and 4G of Ram. The performance criteria of image classification are recall rate, accuracy rate, and Average Precision (AP). The related definitions are as follows,

$$\text{Recall} = \frac{\text{correctly classify image numbers}}{\text{the total image numbers of one category}} \times 100\% \tag{11}$$

$$\text{Precision} = \frac{\text{correctly classify image numbers}}{\text{the total classify image numbers}} \times 100\% \tag{12}$$

$$\text{Average precision} = \frac{\text{sum of precision}}{\text{the total number of image categories}} \tag{13}$$

**Fig. 7.** The sample images of each category

### 4.2 Experimental results

As mentioned in section 2, the hash tables' number L and hash functions' number k are two crucial parameters. The larger the L value, the stronger the algorithm's robustness, but the time efficiency decreases, and $k$ value will affect the number of hash buckets. We firstly construct the visual dictionary with different $k$ values to analysis the superiority of weakly supervised E2LSH. **Fig. 8**(a) depicts the relationship between the number of hash buckets and $k$ value. **Fig. 8**(b) depicts the relationship between the average precision of different methods and the dictionary size. From **Fig. 8**(a), it can be seen that the number of weakly supervised E2LSH cluster centers grows when $k$ value increases, the different methods' average precision grows largely when the dictionary size increases from 300 to 1000, but the precision grows slightly when the size exceeds 1000. For compromise we set $k$=8 and after some noise visual words elimination we set the dictionary size to be 1000. Moreover, form **Fig. 8**(b), we can see that the visual dictionary generated by weakly supervised E2LSH have the better performance than AKM when using the same VLM classification method. It can effectively verify the results of **Fig. 3** and **Fig. 4**.

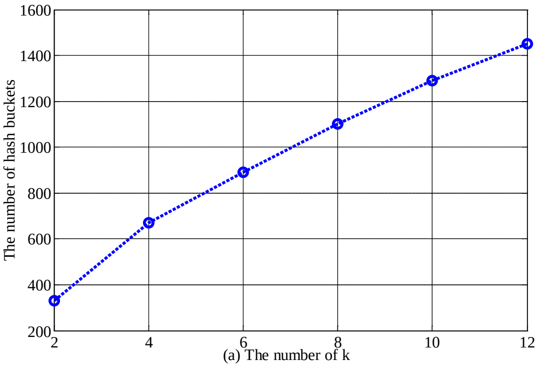


**Fig. 8.** (a) the relationship between the number hash buckets and $k$ value.

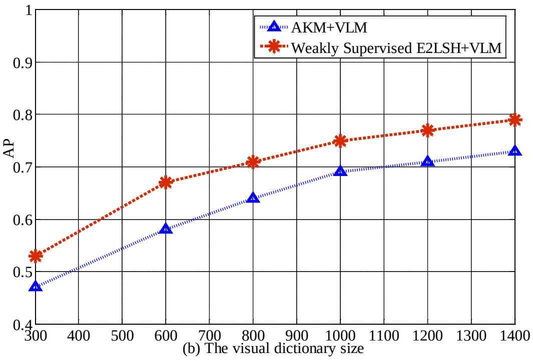


**Fig. 8.** (b) the relationship between the average of Precision and dictionary size

**Fig. 9** depicts the relationship between the average precision of the weakly supervised E2LSH+VLM method and the hash tables' number $L$ when $k=8$. From **Fig. 9** it can be seen that the AP value grows with the $L$ value increases, but as we known that the too lager $L$ value will decrease the efficiency of weakly supervised E2LSH algorithm, so, here we set $L=5$. Besides, we evaluate the time efficiency of weakly supervised E2LSH and the AKM algorithm in generating the visual dictionary. The results are shown in **Fig. 10**. From **Fig. 10**, we can see that the time consumption of the two algorithms increase in a logarithmic form with the increase of dictionary size, but the weakly supervised E2LSH used very fewer time than AKM under the same visual dictionary size. It shows that the weakly supervised E2LSH clustering can improve the time efficiency substantially.

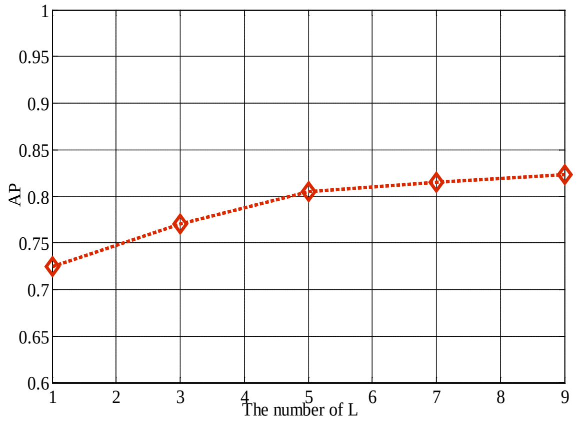


**Fig. 9.** the relationship between the average precision and the hash tables' number $L$

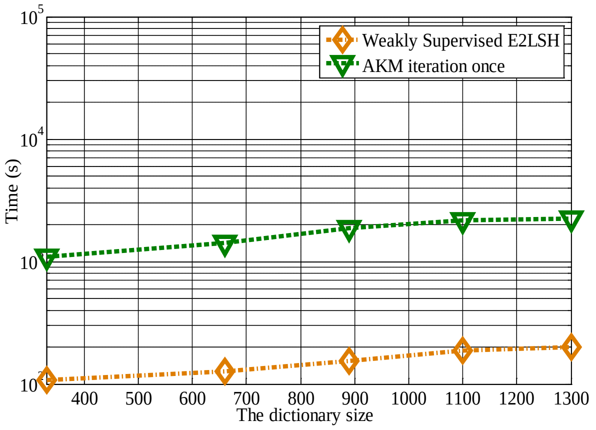
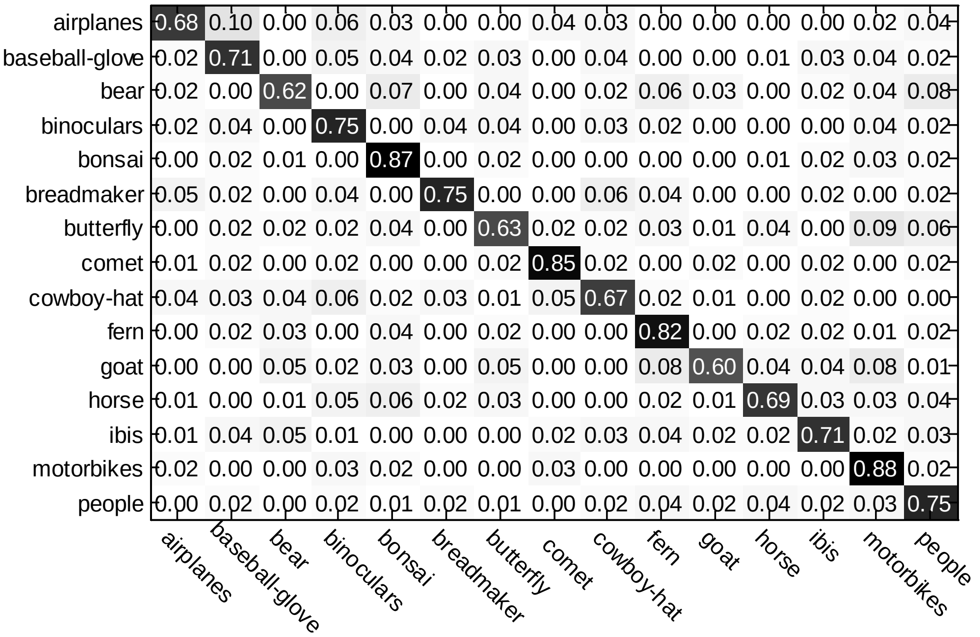


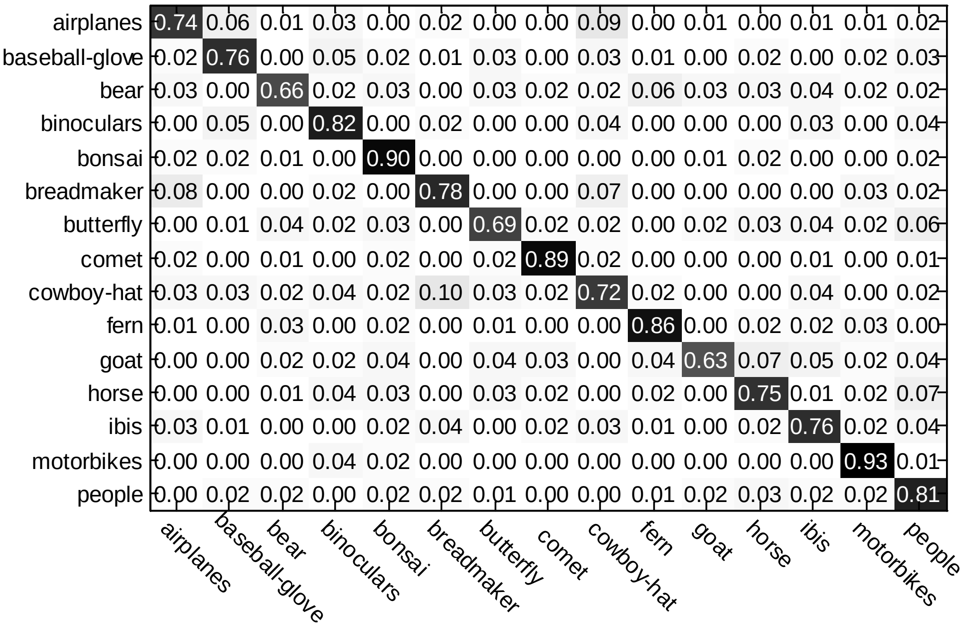
**Fig. 10.** Time efficiency comparison of different clustering algorithms

Then, we adapt traditional one gram visual language model (VLM) and saliency map weighted visual language model ( SMW-VLM) to classify the whole Caltech-256 dataset object categories to verify the effectiveness of saliency map weighted visual language model for object classification under the condition $k=8$， $L=5$, where the dictionary size is 20K after visual words elimination. **Fig. 11** shows the confusion matrix of traditional one-gram visual language model method, and **Fig. 12** depicts the confusion matrix of the saliency map weighted visual language model. From **Fig. 11** and F **Fig. 12**, we can conclude that the saliency map weighted visual language model can improve the recall rate of all the 8 objects classification efficiently. It means that the saliency map weighted visual language model can capture the spatial information effectively as well as overcome the adverse impact of image background noise. However, the performance improvement is slightly different for different objects due to the difference of the training data.



| | airplanes | baseball-glove | bear | binoculars | bonsai | breadmaker | butterfly | comet | cowboy-hat | fern | goat | horse | ibis | motorbikes | people |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| airplanes | 0.68 | 0.10 | 0.00 | 0.06 | 0.03 | 0.00 | 0.00 | 0.04 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.04 |
| baseball-glove | 0.02 | 0.71 | 0.00 | 0.05 | 0.04 | 0.02 | 0.03 | 0.00 | 0.04 | 0.00 | 0.00 | 0.01 | 0.03 | 0.04 | 0.02 |
| bear | 0.02 | 0.00 | 0.62 | 0.00 | 0.07 | 0.00 | 0.04 | 0.00 | 0.02 | 0.06 | 0.03 | 0.00 | 0.02 | 0.04 | 0.08 |
| binoculars | 0.02 | 0.04 | 0.00 | 0.75 | 0.00 | 0.04 | 0.04 | 0.00 | 0.03 | 0.02 | 0.00 | 0.00 | 0.00 | 0.04 | 0.02 |
| bonsai | 0.00 | 0.02 | 0.01 | 0.00 | 0.87 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.03 | 0.02 |
| breadmaker | 0.05 | 0.02 | 0.00 | 0.04 | 0.00 | 0.75 | 0.00 | 0.00 | 0.06 | 0.04 | 0.00 | 0.00 | 0.02 | 0.00 | 0.02 |
| butterfly | 0.00 | 0.02 | 0.02 | 0.02 | 0.04 | 0.00 | 0.63 | 0.02 | 0.02 | 0.03 | 0.01 | 0.04 | 0.00 | 0.09 | 0.06 |
| comet | 0.01 | 0.02 | 0.00 | 0.02 | 0.00 | 0.00 | 0.02 | 0.85 | 0.02 | 0.00 | 0.02 | 0.00 | 0.02 | 0.00 | 0.02 |
| cowboy-hat | 0.04 | 0.03 | 0.04 | 0.06 | 0.02 | 0.03 | 0.01 | 0.05 | 0.67 | 0.02 | 0.01 | 0.00 | 0.02 | 0.00 | 0.00 |
| fern | 0.00 | 0.02 | 0.03 | 0.00 | 0.04 | 0.00 | 0.02 | 0.00 | 0.00 | 0.82 | 0.00 | 0.02 | 0.02 | 0.01 | 0.02 |
| goat | 0.00 | 0.00 | 0.05 | 0.02 | 0.03 | 0.00 | 0.05 | 0.00 | 0.00 | 0.08 | 0.60 | 0.04 | 0.04 | 0.08 | 0.01 |
| horse | 0.01 | 0.00 | 0.01 | 0.05 | 0.06 | 0.02 | 0.03 | 0.00 | 0.00 | 0.02 | 0.01 | 0.69 | 0.03 | 0.03 | 0.04 |
| ibis | 0.01 | 0.04 | 0.05 | 0.01 | 0.00 | 0.00 | 0.00 | 0.02 | 0.03 | 0.04 | 0.02 | 0.02 | 0.71 | 0.02 | 0.03 |
| motorbikes | 0.02 | 0.00 | 0.00 | 0.03 | 0.02 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.88 | 0.02 |
| people | 0.00 | 0.02 | 0.00 | 0.02 | 0.01 | 0.02 | 0.01 | 0.00 | 0.02 | 0.04 | 0.02 | 0.04 | 0.02 | 0.03 | 0.75 |

**Fig. 11.** The confusion matrix of the traditional one-gram visual language model



| | airplanes | baseball-glove | bear | binoculars | bonsai | breadmaker | butterfly | comet | cowboy-hat | fern | goat | horse | ibis | motorbikes | people |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| airplanes | 0.74 | 0.06 | 0.01 | 0.03 | 0.00 | 0.02 | 0.00 | 0.00 | 0.09 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.02 |
| baseball-glove | 0.02 | 0.76 | 0.00 | 0.05 | 0.02 | 0.01 | 0.03 | 0.00 | 0.03 | 0.01 | 0.00 | 0.02 | 0.00 | 0.02 | 0.03 |
| bear | 0.03 | 0.00 | 0.66 | 0.02 | 0.03 | 0.00 | 0.03 | 0.02 | 0.02 | 0.06 | 0.03 | 0.03 | 0.04 | 0.02 | 0.02 |
| binoculars | 0.00 | 0.05 | 0.00 | 0.82 | 0.00 | 0.02 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.04 |
| bonsai | 0.02 | 0.02 | 0.01 | 0.00 | 0.90 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.00 | 0.00 | 0.02 |
| breadmaker | 0.08 | 0.00 | 0.00 | 0.02 | 0.00 | 0.78 | 0.00 | 0.00 | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.02 |
| butterfly | 0.00 | 0.01 | 0.04 | 0.02 | 0.03 | 0.00 | 0.69 | 0.02 | 0.02 | 0.00 | 0.02 | 0.03 | 0.04 | 0.02 | 0.06 |
| comet | 0.02 | 0.00 | 0.01 | 0.00 | 0.02 | 0.00 | 0.02 | 0.89 | 0.02 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 |
| cowboy-hat | 0.03 | 0.03 | 0.02 | 0.04 | 0.02 | 0.10 | 0.03 | 0.02 | 0.72 | 0.02 | 0.00 | 0.00 | 0.04 | 0.00 | 0.02 |
| fern | 0.01 | 0.00 | 0.03 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 | 0.00 | 0.86 | 0.00 | 0.02 | 0.02 | 0.03 | 0.00 |
| goat | 0.00 | 0.00 | 0.02 | 0.02 | 0.04 | 0.00 | 0.04 | 0.03 | 0.00 | 0.04 | 0.63 | 0.07 | 0.05 | 0.02 | 0.04 |
| horse | 0.00 | 0.00 | 0.01 | 0.04 | 0.03 | 0.00 | 0.03 | 0.02 | 0.00 | 0.02 | 0.00 | 0.75 | 0.01 | 0.02 | 0.07 |
| ibis | 0.03 | 0.01 | 0.00 | 0.00 | 0.02 | 0.04 | 0.00 | 0.02 | 0.03 | 0.01 | 0.00 | 0.02 | 0.76 | 0.02 | 0.04 |
| motorbikes | 0.00 | 0.00 | 0.00 | 0.04 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.93 | 0.01 |
| people | 0.00 | 0.02 | 0.02 | 0.00 | 0.02 | 0.02 | 0.01 | 0.00 | 0.00 | 0.01 | 0.02 | 0.03 | 0.02 | 0.02 | 0.81 |

**Fig. 12.** The confusion matrix of the saliency map weighted visual language model

Finally, we do experiment on Pascal Voc 2007 datasets to further verify the effectiveness of our method (WS-E2LSH+SMW-VLM), the dictionary size of are 1000 and 10K respectively.

We compare the average precision of our method (WS-E2LSH+SMW-VLM) with that of AKM+visual language model method[20] and AKM+Geometric Phrase Pooling method (AKM+GPP) [18], and the deep learning method [30] respectively. The average precision of different methods are shown as **Fig. 13**. From **Fig. 13**, we can see that the AP of WS-E2LSH+SMW-VLM is higher than that of AKM+VLM, AKM+GPP and deep learning. This indicates that the generated visual dictionary by weakly supervised E2LSH clustering is superior to AKM in overcoming the problem of synonymy and polysemy of visual words. The saliency map weighted visual language model can capture more spatial information of visual words than the traditional visual language and Geometric Phrase Pooling method (GPP). In a word, our method can improve object classification performance on the whole, but for different object classification tasks, the final performance also depends on the data property, whether the training set is sufficient or not, etc.

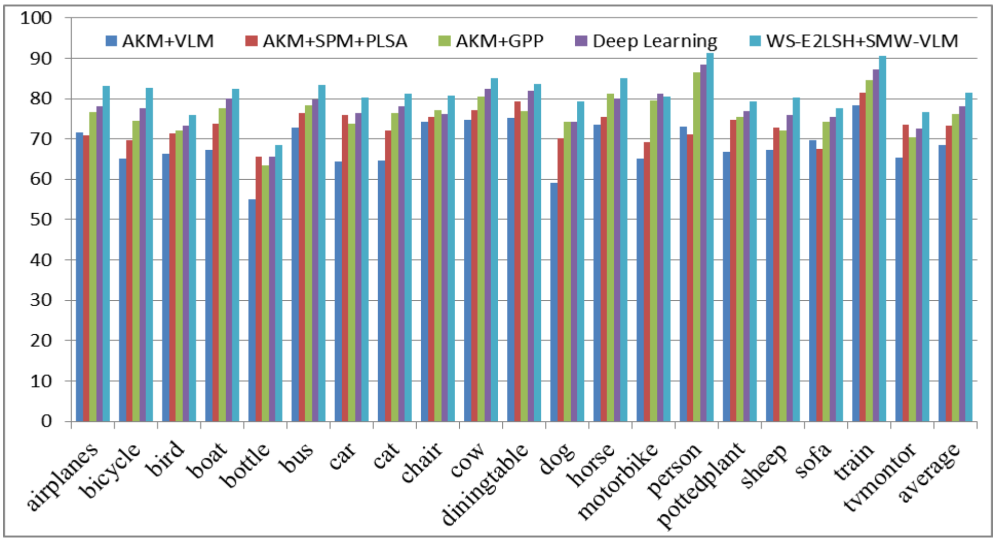


**Fig. 13.** Average precision comparison of different methods on Pascal Voc2007

## 5. Conclusion

In this paper, we propose an object classification method based on weakly supervised E2LSH and saliency map weighting. Firstly, weakly supervised E2LSH is employed to construct a group of weakly randomized visual dictionary, in which a new weakly supervised scheme is proposed to choose out the best hash functions, which can effectively reduce the randomness of traditional E2LSH algorithm and enhance the distinction and extensibility of visual dictionaries. Meanwhile, our method utilizes graph-based visual saliency algorithm to detect the saliency map of images and weight the visual words according to its saliency prior. Finally, saliency map weighted visual language model is carried out to accomplish object classification. The experiment results on Caltech-256 and Pascal Voc 2007 indicate that the weakly supervised E2LSH can overcome the low time efficiency of K-Means clustering algorithm as well as the problem of synonymy and polysemy of visual words. The saliency map weighted visual language model also can capture more spatial information of visual words than the state-of-the-art methods. Despite the excellent accuracy gain we have obtained, there is still one open problem in our framework. It is that how to make the distance metric much closer to the real semantic space through distance metric learning.

## References

[1] J. Sivic, A. Zisserman. “Video Google: a text retrieval approach to object matching in videos,” in *Proc. of 9th IEEE International Conference on Computer Vision*, pp. 1470-1477, October 13-16, 2003. Article (CrossRef Link).

[2] H. Jegou, M. Douze, C. Schmid. “Packing bag-of features,” in *Proc. of IEEE 12th International Conference on Computer Vision*, pp. 2357-2364, September 29-October 2, 2009. Article (CrossRef Link).

[3] Y. Z. Chen, A. Dick, X. Li, et al. “Spatially aware feature selection and weighting for object retrieval,” *Image and Vision Computing*, vol. 31, no. 6, pp. 935–948, December, 2013. Article (CrossRef Link).

[4] J. Y. Wang, H. Bensmail, X. Gao. “Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification,” *Pattern Recognition*, vol. 46, no. 3, pp. 3249-3255, June, 2013. Article (CrossRef Link).

[5] O. A. B. Penatti, F. B. Silva, Eduardo Valle, et al. “Visual word spatial arrangement for image retrieval and classification,” *Pattern Recognition*, vol. 47, no. 1, pp. 705-720, June, 2014. Article (CrossRef Link).

[6] D. G. Lowe. “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, April, 2004. Article (CrossRef Link).

[7] J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders, et al. “Visual word ambiguity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no. 32, pp. 1271-1283, July, 2010. Article (CrossRef Link).

[8] Raphaël Marée, Philippe Denis, Louis Wehenkel, et al. “Incremental indexing and distributed image search using shared randomized dictionaries,” in *Proc. of* MIR 2010, pp. 91-100, May 05-07, 2010. Article (CrossRef Link).

[9] D. Nister, H. Stewenius. Scalable recognition with a vocabulary tree[C], in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2161-2168June . 17-22, 2006. Article (CrossRef Link).

[10] J. Philbin, O. Chum, M. Isard, et a1. “Object retrieval with large vocabularies and fast spatial matching,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, June 17-22, 2007. Article (CrossRef Link).

[11] R. J. Zhang, F.S Wei, B. C. Li. “E2LSH based Multiple Kernel Learning Approach for Object Detection,” *Neurocomputing*, vol. 124, no. 1, pp. 105-110, March, 2014. Article (CrossRef Link).

[12] Q. Zheng, W. Gao. “Constructing visual phrases for effective and efficient object-based image retrieval,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 5, no. 1, pp. 1-19, May, 2008. Article (CrossRef Link).

[13] T. Chen, K. H. Yap and D.J. Zhang. “Discriminative soft bag-of-visual phrase for mobile landmark recognition,” *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 612-622. April, 2014. Article (CrossRef Link).

[14] J. Philbin, O. Chum, M. Isard, et al. “Lost in quantization: Improving particular object retrieval in large scale image databases,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8. June 23-28. 2008. Article (CrossRef Link).

[15] W. Jing-yan, L. Yong-ping, Z. Ying, et a1. "Bag-of-features based medical image retrieval via multiple assignment and visual words weighting," *IEEE Transactions on Medical Imaging*, vol. 30, no. 11, pp. 1996-2011, November, 2011. Article (CrossRef Link).

[16] S. Lazebnik, C. Schmid, J. Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169-2178. October 21-26. 2006. Article (CrossRef Link).

[17] G. Sharma, F. Jurie. "Learning discriminative spatial representation for image classification," in *Proc. of the 22nd British Machine Vision Conference*, pp. 1-11. July 08-11, 2011. Article (CrossRef Link).

[18] L. Xie, Q. Tian, B. Zhang. "Spatial Pooling of Heterogeneous Features for Image Classification," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 1994-2008, May, 2014. Article (CrossRef Link).

[19] Wu Lei, Li Ming, Li Z, et al. "Visual language modeling for image classification," in *Proc. of the International Workshop on Workshop on Multimedia Information Retrieval*. pp. 115-124. June14-17, 2007. Article (CrossRef Link).

[20] Wu Lei, Hu Y, Li M, et al. "Scale-Invariant visual language modeling for object categorization," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 286-294, February, 2009. Article (CrossRef Link).

[21] S. Nakamoto and T. Toriu. "Combination way of local properties, classifiers and saliency in bag-of-keypoints approach for generic object recognition," *International Journal of Computer Science and Network Security*, vol. 11, no. 1, pp. 35-42, July, 2011. Article (CrossRef Link).

[22] M. Datar, N. Immorlica, P. Indyk, V.S. Mirrokni. "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. of the 20th Annual Symposium on Computational Geometry*, pp. 253-262, October 21-25, 2004. Article (CrossRef Link).

[23] M. Slaney, M. Casey, 'Locality-sensitive hashing for finding nearest neighbors," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 128-131, March, 2008. Article (CrossRef Link).

[24] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency[C], in *Proc. of Advances in Neural Information Processing Systems*, pp. 545–552, November 12-15, 2007. Article (CrossRef Link).

[25] L. Breiman. "Random forests," http://www.stat.berkeley.edu/~breiman/RandomForests/ 2014. 07.

[26] L. Itti, C. Koch, and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November, 1998. Article (CrossRef Link).

[27] B. Geng, L. Yang, and C. Xu. "A study of language model for image retrieval," In: *Proc. of IEEE International Conference on Data Mining Workshops*, pp. 158-163, December 6-6, 2009. Article (CrossRef Link).

[28] F.F. Li, R. Fergus, 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, no. 1, pp. 59-70, Augest, 2005. Article (CrossRef Link).

[29] M. Everingham, L. Van Gool, C. K. I. Williams, et al. "The PASCAL Visual Object Classes Challenge Results,"http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/results/index.shtml, 08. 2014.

[30] S. hui, L. Zhenbao, Han Junwei et al. "Learning High-Level Feature by Deep Belief Networks for 3-D Model Retrieval and Recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2154-2167, December, 2014. Article (CrossRef Link).

**Yongwei Zhao** received his B.S. in Electronic Information Engineering from Shandong University, Jinan, China, in 2009, and received his M.S. in China National Digital Switching System Engineering and Technological R&D Center in 2012, and is currently pursuing the Ph.D. degree. His research interests include image processing and classification.

**Bicheng Li** received his M.S. and Ph.D. degrees in China National Digital Switching System Engineering and Technological R&D Center in 1995 and 1998, respectively. He is currently a Professor with Department of Information Science. His research interests include text processing, image processing and pattern recognition.

**Xin Liu** received his BS. degrees in Signal and Information Processing from China National Digital Switching System Engineering and Technological R&D Center in 2013. His research interests include image processing and network security.

**Shengcai Ke** received his BS. degrees in Signal and Information Processing from Zhengzhou Information Science and Technology Institute in 2013, and is currently pursuing the M.S. degree in China National Digital Switching System Engineering and Technological R&D Center. His research interests include image processing.