www.s2is.org
International Journal on Smart Sensing and Intelligent Systems
ISSN 1178-5608

A SCENE RECOGNITION ALGORITHM BASED ON MULTI-INSTANCE LEARNING

[1]Tao Wang , [2] Wenqing Chen and [3]Bailing Wang
[1]Department of Computer Science and Technology,Shaoxing University
[2]College of Information Engineering, Shaoxing Vocational & Technical College ,
[3]School of Computer Science and Technology,Harbin Institute of Technology
Emails: taohit@usx.edu.cn

Submitted: May 17, 2014 ***Accepted: Oct. 12, 2014*** ***Published: Dec. 1, 2014***

Abstract- In Bag of Words image presentation model, visual words are generated by unsupervised clustering, which leaves out the spatial relations between words and results in such shorting comings as limited semantic description and weak discrimination. To solve this problem, we propose to substitute visual words by visual phrases in this article. Visual phrases built according to spatial relations between words are semantic distrainable, and they can improve the accuracy of Bag of Words model. Considering the traditional classification method based on Bag of Words model is vulnerable to the background, block and scalar variance of an image, we propose in this article a multiple visual words learning method for image classification, which is based on the concept of visual phrases combined with Multiple Instance Learning. The final classification model is able to show the spatial features of image classes. Experiments performed on standard image testing sets, Caltech 101 and Scene 15, show the satisfying performance of this algorithm.

Index terms: Image Classification, Multiple Kernel Learning, Bag of Visual Words, Spatial Pyramid Matching

INTRODUCTION

The eyes are the windows of the mind, and humans achieve more information from vision than through other approaches, thus vision is the most important sensation of humans. It was shown by statistics that 80% of all information we obtain came from vision. Visual information is a kind of multimedia information, and it is popular due to such advantages as vividness in expression and abundance in content. We can achieve a lot of important knowledge from visual information and our life is thus greatly improved. Meanwhile, with the development of IT and multimedia technology, large information source libraries such as the Internet come into being. In particular, since the late 1990s, the thrive of digital cameras and the development of the Internet have made more and more multimedia information frequently appear in our daily life. Therein, information mainly conveyed by images, audios and videos has become the mainstream of information service and exchange and plays important roles in research, daily life, communication and so on. Images have descriptive ability and abundant content beyond the capacity of texts, so people prefer to transmit and achieve information via images. An image contains a large amount of information, and the storage of images in the Internet increases astonishingly. Related researches show that currently the amount of data computers all over the world generate and store doubles every other month. There was a research predicted from databases which were five years before that the amount of information in the Internet would increase exponentially. In addition, certain materials show that this pattern will be followed for a long time. Therefore, besides convenience, the mass data has brought many abstruse problems as well: so much visual information has exceeded the receptive capacity of us and the large amount of data have made it harder and harder to process and achieve useful information. In the information explosive ocean, we find it difficult to find required information efficiently. Users may spend large amount of time and energy to find related information, which only turns out unsatisfying. Then, how to pick out the truly useful information in mass data? Visual Intelligence, i.e. Computer Vision, is an important component of Artificial Intelligence. It is a discipline working on how to make a machine ‘see’ and can help us with information searching in mass data. In the process of searching, the first step is to obtain the image information, then the content of the image is analyzed and comprehended and related knowledge is obtain via the description of the image. Image analysis and comprehension is the primary task during image information analysis in computers and an important part of Visual Intelligence.

Meanwhile, the classification of visual objects is critical in image analysis and comprehension. Thus, in the face of increasing mass of images, the automatic classification and extraction of visual objects have become more and more urgent. Visual object classification is to automatically perform object classification on an image or to judge whether an image belongs to a certain class as well as locate and extract target objects from sequential images. It is a hot and difficult problem in Computer Vision and Pattern Recognition and critical to image comprehension and retrieval [1]. For human beings, vision is a learning process of thinking and perception. Its basic approach is: visual organs receive information from outside as sensors, and then they transmit the received stimulus to the brains, which generates a vivid semantic description for visual objects after processing the information. Any visual object has its features and the description of the features is an important prerequisite of visual object classification [2]. Similar to visual information process via human eyes and brains, visual object classification performs learning and sorting on semantic concepts by the features of bottom-layer visual objects and then builds complicated object classification model. The basic idea is: first describe the visual objects and build a model, then adapt the model based on the class of visual objects via Machine Learning, and finally use the adapted model to classify unknown objects. The description of visual objects is to find the objects and describe them via their features. This process mainly faces two challenges. The first one resides in the features of visual objects. Up to now there have been a lot of researches on image features for classification at home and abroad, in which the contents of images are presented by objective visual features such as colors, textures, shapes, spatial relations and so on. However, computers are not capable of high-level semantic description like human being, and the low-level features they can describe are lagging far behind the high-level semantic features humans can understand, i.e. there is a huge semantic gap between low-level visual features in images and semantic description. The existence of semantic gap disables computers to describe visual objects effectively and brings challenges to visual object classification. The second challenge is the difficulty in image object extraction due to visual angle variance, brightness variance, scale variance, object transformation, partial blocking, complicated background and variance within the same class of objects. In addition, the continuing increase of image data makes traditional manual extraction extremely infeasible. All the above influences also add to the semantic gap. Therein, brightness variance is caused by the change of light, scale variance by the rotation of objects and change of shooting angle, and variance within the same class by different postures and appearances of the same object.

The traditional method of visual object classification is to manually label the objects in images and extract features as well as performs classification according to the labels. Its short coming is that it costs an enormous amount of time and effort to work on large or even mass amount of image data and it is also prone to generate errors that we cannot accept. Moreover, in this information explosion era, new images emerge ceaselessly, which makes the instant update of manual labels impossible. So it is infeasible to label all images once for all. In addition, manual labeling causes divergence in labels. An image contains abundant information, and different person have different points of view to the image. For the same semantic information, different person give out different descriptive words. Recently researchers have proposed a series of unsupervised algorithms for visual object recognition. All images are not labeled, so these algorithms bring the highest ambiguity. Unsupervised algorithms are efficient dealing with images sets with solid-color scenes and small number of classes. But they fail to yield ideal results when the color of scenes changes a lot and there are many classes.

Recently, Bag of Words image presentation model has drawn more and more attentions in the area of image classification. It is originally a simplified postulated model to process natural language and retrieve information. In this model, a text can be presented as an unordered set of words, regardless of grammar and word sequences. It was primarily applied to text classification (Lewis, 1998). Now, more and more Bag of Words models are widely employed in visual object classification. The basic idea is to extract local features of an image set in the first place, then quantify these features, and finally present the images as a set of several visual words. Essentially the model is divided into two parts: (1) coding process, which substitutes descriptors with expressions more suitable for classification; (2) merging process, which summarizes features after coding in wider ranges. In the coding process, K-Means, GMM, Sparse Coding are generally used. In the merging process, SPM and Spatial-LTM are frequently employed. K-Means is a distance based clustering algorithm, i.e. the closer two objects are, the more similar they are supposed to be, and finally all data can be divided into k clusters. Sparse Coding is popular these years. It uses as few as possible data points for coding, provided that sparse effectiveness is maintained. However, some problems exist in these popular coding methods, mainly: (1) they leave out the spatial relations among local feature, which makes the generated visual words weak in discrimination and spatial description; (2) they can only judge the existence of an object, but not its location, which causes weak labeling of image.

To address the above problems, i.e. the time-consuming effort in manual labeling and the unsuitability of unsupervised methods in real image classification, Multiple Instance Learning (MIL) is applied to classification of weakly labeled images (i.e. only the existence of objects is known). In this kind of learning, every image in a set is regarded as a bag. Extract the features in blocks and regions or local as samples of this bag so that each bag corresponds to a sample set. Each bag has a training label, which is positive when the image contains the target object and negative otherwise. Samples do not have training labels. In particular, MIL algorithms are discriminative classifiers rather than generative classifiers, which avoid complicate deduction and have higher classification accuracy. Thus MIL has drawn wide attentions in the area of computer vision. For example, Maron et al [3] employed Diverse Density (DD) algorithm to perform classification on natural scenes. DD-SVM model uses DD algorithm to challenge the original model and uses SVM in the original model to classify bags. In most MIL algorithms, it is assumed that there is at least one positive sample in a positive bag and all samples in negative bags are negative. Although this assumption is effective in drug activity prediction, it faces great limit in other applications, especially in images. On the other hand, Chen et al did not make such assumption in MILES algorithm; rather, they mapped bags to feature spaces and used 1-norm SVM as the bag classifier. This process transforms an MIL problem into a standard supervised learning problem, which is more robust against labeling uncertainty.

Considering the great achievement Bag of Words model has made in image classification, the research in this article is apt to include its idea in MIL. However, this model leaves out the spatial relations among local features and makes the generated visual words weak in discrimination and spatial description. Many articles have probed into the spatial relations of visual words [4-6]. The current tendency of the development of this model is mainly represented by intermediate features between image features and bottom ones. For example, intermediate presentation generated via image dissection, Spatial Pyramid and Visual Phrase or Visual Synset proposed via the geometric similarity and spatial relations of local features. Some recent works[7-9] built high order image features based on visual words and showed good results in experiments [7] selected important features and screened low order features for high order features using discrete AdaBoost algorithm. Compared to traditional method of enumeration, this method avoids a large amount of calculation. However, this method is complicated that every time new features are to be selected, AdaBoost algorithm is used to classify training data. [9] proposed a method that searched for meaningful visual phrase vocabulary. First of all, each visual word are combined with its K-nearest neighbors to form visual word group. Then the redundancy in visual phrases is dealt with via mode summary and top-to-bottom update method. Visual Synset [8] consists of visual phrases. It is compared with traditional Bag of Words model and proved to be better than the latter. However, similar to the method proposed by [9], K-nearest Neighbor is also used to construct visual phrases in this method.

MIL can address problems in weak labeling effectively. It is a discriminative training model and brings high-precision classification. But some problems exist in the large numbers of MIL algorithms, so they cannot be applied to visual object classification directory. To overcome the shortcomings in Bag of Words model that unsupervised clustering during generation of visual words results in lack of spatial information, limited sematic description and weak discrimination, we substitute visual words with visual phrases that is more semantically discriminative and contains spatial relations. This improves the accuracy of Bag of Words model. In addition, considering the traditional classification method based on Bag of Words model is vulnerable to the background, block and scalar variance of an image, we propose in this article a multiple visual words learning method for image classification, which is based on the concept of visual phrases combined with Multiple Instance Learning. The final classification model is able to show the spatial features of image classes. Experiments are performed on standard image testing sets, Caltech 101 and Scene 15 [10].

RELATED WORKS

In this chapter, we will introduce the framework of visual object classification based on SPM algorithm [11]. Then we will introduce SIFT descriptor [12], which is a frequent used local feature in this article. In the end, an important algorithm in visual object classification, Multiple Instance Learning (MIL), is presented.

A. Framework of visual object classification based on SPM algorithm

To address the problem mentioned in the first chapter that Bag of Words model neglects spatial relations, Lazebnik et al proposed Spatial Pyramid Matching (SPM) method in 2006, which is applicable to natural scene classification. In this method, an image is intersected into several sub-blocks and the local feature histogram of each sub-block is made. First of all, suppose ***X*** and ***Y*** are two sets of feature vectors in a d-dimensional space. Construct grids in the level of 0,…,*L* that there are 2*l* grids when the level is *l*. Assign H^l_X and H^l_Y respectively as the histograms of ***X*** and ***Y*** in Level *l*, then $H^l_X(i)$ and $H^l_Y(i)$ are the numbers of ***X*** and ***Y*** in the i^th grid. The matching function of Level *l* is defined as:

$$L(H_X^l, H_Y^l) = \sum_{i=1}^{D} \min(H_X^l(i), H_Y^l(i)) \qquad (1)$$

To combine all sub-blocks, the pyramid matching kernel is defined as:

$$\begin{aligned} k^L(X,Y) &= I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (I^L - I^{L+1}) \\ &= \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^L \end{aligned} \qquad (2)$$

Pyramid with three levels is shown as figure 1:

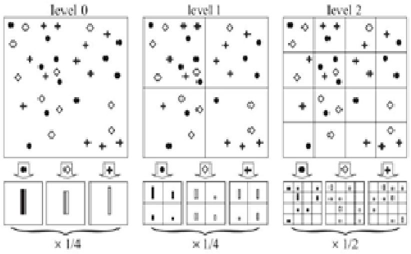

Figure 1 Spatial pyramid matching with three Levels

SPM algorithm has an everlasting influence on object recognition. It is shown that this algorithm in a way includes the spatial relations of local features. As depicted in Figure 1, the weight of every level is fixed when the number of levels is 3. Although fixed weights simplifies the learning process, it is not proper for the whole, because the weight of region containing target object is surely increased.

B. A brief introduction of SIFT

SIFT descriptor (Lowe, 2004) is a local features, based on scale space and invariant to scaling, rotation and even affine transformation. It is forwarded by Lowe in 2004 on the basis of present feature detection methods based on invariant techniques. Its full name is Scale Invariant Feature Transform. Due to its robustness, SIFT is an important feature throughout this article. In the last section, the local feature used in SPM is a SIFT descriptor. This section will briefly introduce its strong points and generation. SIFT descriptor has the following advantages:

(1) SIFT is a local feature, which is invariant to rotation, scaling and change of light and stable in a certain extent of changes in visual angle, affine transformation and noise.

(2) It contains abundant information and has great specificity, so it is applicable to fast and accurate matching among mass feature data.

(3) It has a large quantity that even a few objects can generate a number of SIFT features.

(4) It is high-speed, which makes extraction convenient and fast and satisfies the requirement of real-time.

(5) It is extensible and easy to combine with other feature vectors.

Before the extraction of SIFT, an image is presented in multiple scales. Gauss Kernel Convolution is the only transformation kernel to realize scalar changes of images, and it is the only linear kernel:

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma}e^{-(x^2+y^2)/2} \qquad (3)$$

Therefore, the presentation of a two-dimensional image in different scales can be obtained from the convolution of the image and Gauss Kernel:

$$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y) \qquad (4)$$

Therein, σ is called Scale Space Factor. The smaller σ is, the less smooth the image is and the smaller the scale is. Large scale corresponds to general features of the image, and small scale to detailed features.

Firstly, SIFT algorithm detects features in the scale space and confirms the location and scale of key points. Then, it sets the direction of gradient as the direction of the point. Thus the scale and direction invariance of the operator are realized. Key points are the extreme values in both the two-dimensional space of the image and DoG scale space. DoG operator is calculated as:

$$D(x,y,\sigma)=L(x,y,k\sigma)-L(x,y,\sigma) \qquad (5)$$

Find out all local extremes in DoG space and group them as candidate key points. Then delete the low-contrast key points and unstable skirt response points (for DoG algorithm will generate strong skirt responses) to strengthen the stability and resistance to noise of matching. After obtaining the candidate key points, accurately determine the locations of key points using surrounding data points. Therein, the locations and scales of key points are determined via fitting three dimensional quadratic functions.

Calculate the direction of a key point via the distribution of the direction of gradient of its neighboring pixels and set the direction parameters for each key point to ensure the rotation invariance of the operator. The algorithm samples in the window centered at the key point and calculate the direction of gradient in the neighboring area via histogram. The range of gradient histogram is 0~360 degree, in which each column spans 10 degrees and there are 36 columns. The peak value of the histogram is the principal direction of the neighboring gradient, i.e. the direction of this key point. In this histogram, if a peak value that is 80% of the principal peak exits, the corresponding direction is the auxiliary direction of this key point. A key point may be assigned to several directions (one principal and more than one auxiliary), which can increase the robustness of SIFT feature.

Up to now, the detection of key points is completed. Each key point has three parameters: location, scale and direction. Thus an SIFT feature region can be determined. A SIFT feature vector contains the information of direction in neighborhood and therefore enhances the anti-noise ability as well as provides a good fault tolerance for matching with deviated locations. Get an 8×8 window centered on the key point. First of all, rotate the axis to the direction of key point. Then assign a weight to each direction using Gauss Smooth Filter and calculate the direction histogram of every sub-window. Generally, each key point is described by 4×4 seed points. Thus, 128 data points are generated for one key point. Now SIFT vector is free from the influence of geometric transformations such as scale changes and rotation. Normalize the length of the feature vector, and the influence of light is eliminated.

C. Multiple Instance Learning

Due to its merits in processing weakly label data and performing high-precision classification, Multiple Instance Learning (MIL) is extensively used in Computer Vision. Since the 1990s, MIL has been regarded as the most promising machine learning method. It was proposed by Dietterich et al in the 1990s to judge whether a medical molecule is a Musk molecule.

MIL is the fourth learning framework of machine learning, the other three are Supervised Learning, Unsupervised Learning and Reinforcement Learning. Supervise Learning is the most traditional one. It performs learning on labeled training samples and predicts the non-training data as accurately as possible. In such learning framework, all training samples are labeled to ensure the lowest ambiguity. However, this method costs a lot of time and effort. Unsupervised Learning can perform learning directly on unlabeled training samples and discover the hidden data structure during the learning process. None of the samples is labeled, so this model brings the highest ambiguity. Compared to Supervised Learning in which all samples are labeled, MIL does not have the concept label. Compared to Unsupervised Learning where none of the samples is unlabeled, MIL has labeled bag. Unlike Reinforcement Learning, MIL does not have the concept of time delay. More importantly, in all learning frameworks, a sample is an instance, i.e. there is a one-to-one relation between samples and instances. The ambiguity in Supervised Learning, Unsupervised Learning and Reinforcement Learning is different from that in MIL, which makes the former three unsuitable to deal with such problems. MIL has unique properties and promising application prospect and it is a green area in Machine Learning, so it is popular at home and abroad and regarded as a new learning framework. It has drawn the attention of many researchers, and a lot of new theories and striking application results emerge in a short time.

After the proposal of MIL by Dietterich et al, a lot of effort has been put to this learning framework in the area of Machine Learning, so researches on MIL are very active. In 1998, O. Maron et al put forward Diverse Density (DD) algorithm. Generally, every instance contained in a candidate molecule is presented as a feature vector in an n-dimensional space. The point corresponding to the feature vector will form a trajectory in the n-dimensional space with the change of the molecule’s shape. In a positive bag, there is at least one positive instance, i.e. along the trajectory of a positive bag there is at least one point whose corresponding molecular shape fits the shape of the target molecule. While in a negative bag, there is no such positive instance and thus there is no such point along the trajectory.

In this article, we use an MIL algorithm developed from DD algorithm and MILES algorithm. In 1998, O. Maron and T. Lozano-Perez proposed DD based Multi-instance algorithm. The algorithm aims to find the point with the largest diverse density, i.e. target feature t, which is defined as the following: for a point in the feature space, the more positive bags appears around it and the farther negative instances are from it, the larger its diverse density is.

Set $\mathbf{B}_{\mathrm{i}}^{+}$ as the i^th^ positive bag, B_{ij}^{+} as the j^th^ instance in $\mathbf{B}_{\mathrm{i}}^{+}$ and B_{ijk}^{+} the value of the k^th^ feature of B_{ij}^{+}. Similarly, we can define $\mathbf{B}_{\mathrm{i}}^{-}$, B_{ij}^{-} and B_{ijk}^{-} in negative bags. Let *t* present the point with the largest diverse density, then *t* can be determined via maximization of $\Pr(t|\mathbf{B}_1^{+},\ldots,\mathbf{B}_{l+}^{+},\mathbf{B}_1^{-},\ldots,\mathbf{B}_{l-}^{-})$. Suppose the bags are independent on each other, then the formula can be expressed according to Bayes:

$$\arg\max_x \prod_{i=1}^{l^+} \Pr(x=t\,|\,\mathbf{B}_i^{+}) \prod_{i=1}^{l^-} \Pr(x=t\,|\,\mathbf{B}_i^{-}) \,(6)$$

Noise-or model can be employed to specialize the product in the above formula:

$$\Pr(x = t \mid \mathbf{B}_i^+) \propto \max \exp(-\frac{\left\|B_{ij}^+ - t\right\|^2}{\sigma^2}) \qquad (7)$$

$$\Pr(x = t \mid \mathbf{B}_i^-) \propto 1 - \max \exp(-\frac{\left\|B_{ij}^- - t\right\|^2}{\sigma^2}) \qquad (8)$$

In this formula, σ is the parameter to adjust the scale. Since there are multiple local extremes in the diverse density space, the above formula can be solved by Expectation Maximization (EM). Each positive instance is selected as the starting point for searching, thus when t is located via Gradient Descent algorithm, selection can be performed on each property of instance features. However, due to the problem in calculation amount, DD algorithm several rounds of searching. Y. X Chen et al described DD algorithm from the aspect of feature selection and further assumed every instance in the training bag as candidate, thus they put forward MILES (Multiple Instance Learning via Embedded Instance Selection) algorithm. In MILES, all instances are selected as candidate feature, i.e. C={x^k}, k=1,2,…,n, n is the number of instances in the bag. Define the feature vector of Bag Bi as:

$$m(\mathbf{B}_i) = [s(x^1, \mathbf{B}_i), s(x^2, \mathbf{B}_i), ..., s(x^n, \mathbf{B}_i)]^T \qquad (9)$$

The similarity between Instance x^k and Bag $\mathbf{B}^i$ is measured by:

$$s(x^k, \mathbf{B}_i) = \max{}_j \exp(-\frac{\left\|x_{ij} - x^k\right\|^2}{\sigma^2}) \qquad (10)$$

Given l^+ positive bags and l^- negative bags, all the similarity measurements of between bags and instances form the following matrix:

$$[m_1^+, ..., m_{l^+}^+, m_1^-, ..., m_{l^-}^-] = \begin{bmatrix} s(x^1, \mathbf{B}_1^+) & ... & s(x^1, \mathbf{B}_{l^-}^-) \\ s(x^2, \mathbf{B}_1^+) & ... & s(x^2, \mathbf{B}_{l^-}^-) \\ ... & ... & ... \\ s(x^n, \mathbf{B}_1^+) & ... & s(x^n, \mathbf{B}_{l^-}^-) \end{bmatrix} (11)$$

Each row of the matrix is the similarities between an instance and every bag. In this instance-to-bag mapping space, the MIL problem is converted to a Supervised Learning problem. Perform learning on the above matrix via SVM or 1-Norm SVM as follows:

$$
\begin{aligned}
&\min_{w,b,\xi,\eta} \lambda\sum_{k=1}^{n}| w_k | + c_1\sum_{i=1}^{l^+}| \xi_i | + c_2\sum_{j=1}^{l^-}| \eta_j | \\
&s.t.(w^T m_i^T + b) + \xi_i \geq 1, \qquad\qquad (12) \\
&-(w^T m_j^T + b) + \eta_j \geq 1, \\
&\xi_i, \eta_j \geq 0, i = 1,\ldots,l^+, j =,\ldots,l^-
\end{aligned}
$$

c_1 and c_2 are penalty factors used to address the problem of unequal numbers of positive and negative bags. For a new bag **B'**, its judgment can be determined by:

$$f(\mathbf{B}') = \sum_{k=1,2,\ldots,n} w_k s(x^k, \mathbf{B}') + b \qquad (13)$$

MIL was originally used to predict drug activity. With the on-going of MIL researches, its application area is expanding ceaselessly. Maron et al were the first to apply DD algorithm to image classification in 1998. In this application, images containing figures were labeled positive, otherwise labeled negative. Several image blocks were sampled as instances of a bag. Yang et al applied MIL to content based image retrieval in 2000. They pointed out that compared to natural scene images, MIL were more suitable for object based image retrieval. In 2002, [13] compared DD with EM-DD on their performance in image retrieval. Andrews et al (2002) and Huang et al (2002) [14] also employed MIL in image retrieval. In 2002, Andrews et al employed SVM (Support Vector Machine) based MIL to text classification. They presented each chunk of text as a bag consisting of overlapping segments, and the instances were segments containing 50 letters. Their work was tested on TREC9 text classification dataset. Zhou et al applied MIL to Web Index Page Recommendation in 2005. They treated the union of all links in an index page as a bag and links as instances of the bag, thus the problem was converted to an MIL problem. Based on this, they proposed Fretcit-kNN algorithm to solve the page recommendation problem.

OBJECT RECOGNITION BASED ON MULTI VISUAL PHASES LEARNING

Image classification is the prerequisite of visual object classification. It aims at automatically classify a set of images according to their sematic contents and it is a hot point and difficult in Computer Vision. In the first chapter, we mentioned that the complicated distribution of images and the influences of light, view angle as well as screening put obstacles to model construction. Therefore, image classification is always a challenging problem in Computer Vision. A large number of researches on the features used in visual object classification have been carried out at home and abroad. Objective features such as colors, textures, shapes and spatial relations of objects are used to describe the content of images. However, a huge semantic gap exists between the bottom-layer features and the semantic expression. Therefore, Machine Learning is needed to perform sorting and learning on semantic concepts according to bottom-layer features and construct complicated classification model.

Considering Bag of Words model and related methods such as pLSA [15] (probabilistic Latent Semantic Analysis), LDA (Latent Dirichlet Allocation) [16] in recent years, its basic idea is as follows: firstly extract bottom-layer features and quantize them to present images in Bag of Words model, then perform object recognition and image classification using some techniques in text analysis and retrieval. Due to its great success in text analysis and retrieval, Bag of Words model has attracted profound attention of researchers. Inspired by Bag of Words model, researchers have started to employ similar models for image description and achieved some positive results. Cao et al (2007) proposed Spatially Coherent Latent Topic model (Spatial-LTM) [17] to make sure the coherence of latent topics in the same block via spatial interdependency. Lazebniket et al (2006) put forward a scene classification algorithm based on approximate global geometric relations, which can divide an image into a pyramid and summarize the local features of sub-blocks. This sort of methods usually deal with local features of images, such as SIFT (Scale Invariant Feature Transformation) (Lowe, 2004) and HOG (Histogram of Gradient) [18], and form visual words via spatial division of features. For example, K-means can be employed to cluster SIFT features and the clusters centers are selected as visual words. For a given image, its local feature can be roughly represented by the nearest visual word. Thus, similar to the application on text analysis, the distribution of local features on different visual words can construct the Bag of Words model of images.

It should be noticed that although the Bag of Words model of images is similar to that of texts, there are still obvious differences. The main problem is that visual words are generally quantified directly from bottom-layer features via unsupervised learning and the unsupervised quantification method always leads to limited discriminative ability of visual words. In addition, this model leaves out the spatial relations of local features, so the descriptive ability is restricted. There Bag of Words model suffers from weak semantic discrimination and spatial description and the generated visual words face the phenomena of polysemy and synonyms. Due to polysemy and synonyms, features describing different objects cannot be distinguished and are matched. For another, image classes are often than not spatial related but Bag of Words model neglects the spatial relations of visual words. Related studies have shown that local regions in an image are interdependent and constructing models using spatial relations of regions can improve the accuracy of classification [8]. For example, in Caltech 101 image set, class-information contains the existence of specific visual objects or regions. Even if in scene related image classification, such as Scene 15 image set, the similarity among the same class is reflected on some local regions. Moreover, the distribution of image contents is complicated and varied, different images have large divergence on the location, scale and orientation of related regions. Lacking this region-related information, present Bag of Words model is obtained through statistics of the whole image. Therefore corresponding classification model is not accurate enough because it is not capable of demonstrating the regional properties of different image classes and vulnerable to image background, the position of a region and scale. To address this problem, people proposed the methods of constructing visual phrases or visual synset. These methods mainly use the spatial coherent relations of visual words to generate high-order visual features.

To address the problems of limited discrimination and description in Bag of Words model as well as the influences of background and screening, we propose Multiple Visual Phrase Learning (MVPL) image classification method. We substitute visual words with visual phrases that are more semantically discriminative and contain spatial relations to improve the accuracy of Bag of Words model. In order to depict regional properties in the final model, we propose in this article a multiple visual words learning method for image classification, which is based on Multiple Instance Learning. The detailed construction process is as follows: first of all, construct semantic discriminative and spatial related visual phrases in place of visual words to improve the accuracy of image presentation. On this basis, to depict regional properties of image classes in the final model, we propose in this article a multiple visual words learning method for image classification, which is based on MIL and uses the visual phrase library as new feature space. To test the function of MVPL proposed in this article, we conducted many experiments on standard testing datasets. Firstly, we compare the influences of visual words and phrases on image classification. Then we compare our method with some present methods, such as SPM and Spatial-LTM, and the comparisons are performed on some standard testing sets such as Caltech 101 and Scene 15 to prove the effectiveness of our method.

A. Visual phrases

Visual words are generated via Unsupervised Learning, so the related Bag of Words model suffers from limited semantic discrimination. In addition, spatial relations of words are not considered during the generation process. For a given dataset, visual words with certain semantic discrimination are generated in association with class-information. Then visually consistent regions are formed via image over-segmentation. Spatially confine visual words via these regions and construct a visual word set, thus the spatial relations of visual words are realized. In the end, generate visual phrases from the visual word sets of all regions via Frequent Item Set Mining. The detailed generation process will be given as follows.

Given a labeled image, we first detect key points using the method proposed by [19-21] and extract SIFT features of the key points. Then, cluster features of images in different classes separately via K-Means. In the end, group all clustering centers to form the class-related visual work vocabulary. Assign every instance according to these centers and give it corresponding label.

A visually consistent region means a region within which the color, texture et al are coherent on some level. Hereby we use Multi-Scale Ncuts proposed by T. Cour et al to intersect images, meanwhile a threshold value is used to filter out small regions. Please notice that the aim to generate visually consistent regions is to analyze the spatial relations of visual words, and it does not depend on the accuracy of image intersection, the reason of which is to ensure that each region is a part of an object. We employ Multi-Scale Ncuts because it algorithm takes image information under different scales into consideration, which is possible for parallel calculation to improve the speed. In addition, the number of intersected regions can be controlled manually. During our experiments, each image is divided into 30 blocks and 30 or so visually consistent regions are generated eventually. Visual phrases are the combination of visual words which appear frequently in visually consistent regions. Here related techniques in data mining are used to generate visual phrases.

An image $\mathbf{X}_\mathrm{i}$ is divided into J visually consistent regions, $\mathrm{R_i}=\{r_\mathrm{j}^\mathrm{i}\}$, j=1,…,J, and the visual words belonging to the j$^{\mathrm{th}}$ region $\mathrm{r_j^i}$ is presented as a set $G_\mathrm{j}^\mathrm{i}=\{w_1,\ldots,w_{\mathrm{Kj}}\}$, in which K_j is the total number of words in r_j^i. For N images, the set containing all visual words is $\mathrm{G}=\{G_\mathrm{i}\}$ (i=1,2,…,N), in which $G_\mathrm{i}=\{G_\mathrm{j}^\mathrm{i}\}$ (j=1,2,…,J), and the set of visual phrases is $\mathrm{VP}=\{vp_\mathrm{k}\}$ (k=1,…,K). A visual phrase vp is generated from G according to the frequency of the coexistence of related visual words. If regard G as a database, then the construction of visual phrases can be converted into the problem of mining frequent items in data mining, i.e. searching for coherent set from candidate sets consisting of visual words. In this work we use FP-growth (Frequent Pattern growth) algorithm to accomplish this task. FP-growth constructs a tree based on the frequency of units after scanning the database, and frequent item mining is performed by traversing this tree structure. The storage requirements and calculation rate of FP-growth make it suitable for large datasets.

B. Multiple visual phrases

This section will propose a Multiple Visual Phrase Learning (MVPL) method based on Multiple Instance Learning (MIL). In MIL, a training sample is a bag consisting of multiple instances. A bag has a label while a instance does not. It is generally assumed that there is at least one positive instance in a positive bag and on positive instance in a negative bag. The process of MIL not only requires related model parameters, but also the confirmation of positive instances in positive bags. As is mentioned before, the classes of images are region related and present marks often only show the existence of certain region, so the image classification can be essentially converted into an MIL problem. We proposed MVPL starting from visual phrase based Bag of Words model.

Given a visual phrase set $C_{vp}=\{p_k|k=1,2,\dots,K\}$ in which $p_k=\{w_p^k|p=1,\dots,P^k\}$, regard C_{vp} as an instance space. Suppose that an image $\mathbf{X}_j$ is divided into J regions $R^i=\{r_1^i, r_2^i, \dots,r_J^i\}$ and visual words in r_j^i form the word set $G_j^i=\{w_1,\dots,W_{Kj}\}$, in which K_j is the number of visual words in r_j^k. Regard $\mathbf{X}_i$ as a bag and G_j^i as an instance in it: $\mathbf{X}_i=\{G_j^i|j=1,2,\dots,J\}$. Define the similarity between the Visual Phrase p_k and Bag $\mathbf{X}_i$ as:.

$$s(p_k,\mathbf{X}_i)=\begin{cases}1, & if \quad p_k\in\mathbf{X}_i\\ 0, & otherwise\end{cases} \qquad (14)$$

In addition, considering the histogram of visual phrases is better for image presentation, the similarity can be defined as:

$$s(p_k,\mathbf{X}_i)=\sum_j h_{ij}(p_k) \qquad (15)$$

In the above formula, $h_{ij}(p_k)$ is the number of visual phrases contained in the jth region of $\mathbf{X}_i$ and Σh_{ij} is the total number of visual phrases contained in the image. Traditional method can be used to solve the problem. Considering the mapping matrix is a sparse one, L1-norm SVM is still used for image classification.

MIL based on the above formula is MVPL, short for Multiple Visual Phrase Learning. Histogram of visual phrases can better distinguish the descriptive information of different classes of images. In addition, visual phrases contain the relations of visual words extracted from images, so they are more discriminative and descriptive than visual words.

EXPERIEMNTS AND ANALYSIS

Image sets used in the experiments are Caltech 101 and Scene 15, which this section will introduce briefly. As introduced in the last chapter, Caltech 101 contains images related to visual objects such as animals, vehicles and flowers. There are 102 image classes and each class contains 31 to 800 images, whose sizes are 300×300 pixels or so. Scene 15 contains indoor scenes (such as offices, kitchens and living room), outdoor scenes (such as the appearances of buildings, cities and streets) and natural scenes (such as dunes, mountains and forests). Each class contains 200 to 400 images, whose average size is 300×250 pixels. Figure 2 shows three randomly picked images from each class.

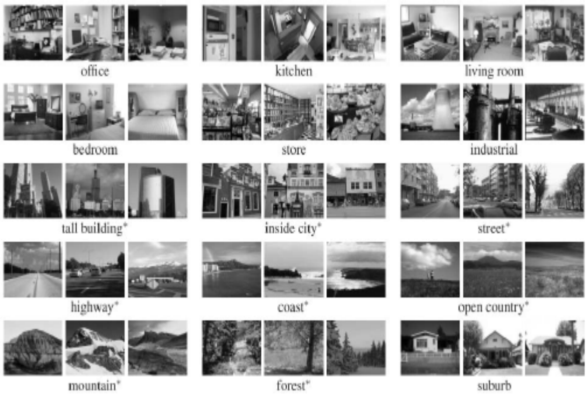

Figure 2 Examples of Scene-15

For a given image set, detect the key points of every image via Harris-Affine algorithm and extract the related SIFT descriptors. Perform clustering on SIFT features belonging to the same class of images via K-Means and generate related visual words, meanwhile construct visual phrases in combination of image intersection. When visual phrase set is given, images can be expressed by related Bag of Words model. In the meantime, we compare our method to some present algorithms such as SPM and Spatial-LTM to show the effectiveness of our method. In the experiments on Caltech 101, we use the same configuration as in the article of Cao et al (2007), i.e. 28 classes of objects are selected, with each class containing at least 60 images. We randomly pick 30 images as the training set and 30 as testing set. The experiment is repeated for 10 times and the average accuracy and the deviation are used to measure the functions of algorithms. For Scene 15, we apply the experimental configuration of SPM, i.e. 100 images of each class are randomly selected as the training set and the remaining as testing set. We not only compare our method with some classical ones, but also compare our method with method using visual words, which is written as MVWL. In MVWL, each visual word is an instance and the related Bag of Words model is the bag B. The similarity is defined as:

$$s(w_k, \mathbf{X}_i) = \begin{cases} 1, & if \quad w_k \in \mathbf{X}_i \\ 0, & otherwise \end{cases} \qquad (16)$$

A Experimental results

Caltech 101 contains 102 image classes such as animals, vehicles, flowers and backgrounds, and each class contains 31 to 800 images, whose sizes are 300×300 pixels or so. To compare our method with Spatial-LTM proposed by Cao et al (2007), we use the same experimental configuration as in their article, i.e. select 28 classes which contain more than 60 classes and randomly select 30 images as training set and 30 as testing set in each class. We observed that these 28 classes contain images with ragged edges such as sunflower, narrow images such as flamingo and images with surrounding such as cup. The diversity of image contents basically reflects the attribute of Caltech 101. The experiment is repeated for 10 times, each time we record the average accuracy of the 28 classes and the average as well as deviation of these 10 experiments are calculated.

Quantize the SIFT features of the 28 classes of training image into visual words with the length of N1 via K-Means, and we obtain 28*N1 visual words eventually. Divide each image into 30 regions via Multi-Scale Ncut and group the visual words from the same region. All images form Set G and extract visual phrases efficiently from G using FP-growth.

First of all, we experiment on the length of visual word set via VM-MIL and the result is shown in Table 1. It is shown that with the increase of visual word length, the recognition accuracy gradually increases, but the amount of calculation also goes up. Therefore, we fix N1 to 200 in the following experiments, thus the length of visual word set is set to 5600.

Table 1 The average classification accuracies of different lengths of visual word set

Length	700	1400	2800	5600
Accuracy	0.423	0.558	0.590	0.697

On the basis of visual word set with the length of 5600, we construct a visual phrase vocabulary whose length is about 40,000 with the visual phrase generation algorithm proposed in this chapter. The comparisons among visual phrases based MIL, Spatial-LTM and SPM are shown in Table 2. It is shown that VW-MIL is already obviously better than Spatial-LTM and SPM. The accuracy of MVWL is almost **% higher than VW-MIL and nearly **% higher than

Spatial-LTM and SPM. The classification performance of visual phrases is even better than visual words.

Table 2 The average accuracies of different methods

Spatial-LTM	68.9%
SPM	67.9%
VW-MIL	70.1%
MVWL	76.31%
MVPL	76.20%

The second experiment is performed on 15 Scene. There are 15 classes of scenes in 15 Scene and each contains 200 to 400 images, whose average size is 300×250 pixels. Scene 15 contains indoor scenes (such as offices, kitchens and living room), outdoor scenes (such as the appearances of buildings, cities and streets), natural scenes (such as dunes, mountains and forests) and artificial scenes (such as highway and suburb), and it is the most thorough scene dataset. The experiment is repeated for 10 times and its configuration is the same as in SPM algorithm (Lazebnik et al, 2006), i.e. randomly pick 100 images from each class as training set and the remaining as testing set.

Quantize the SIFT features of the 15 classes of training image into visual words with the length of N2 via K-Means, and we obtain 15*N1 visual words eventually. Divide each image into 20 regions via Multi-Scale Ncut and group the visual words from the same region. All images form Set G and extract visual phrases efficiently from G using FP-growth.

First of all, we experiment on the length of visual word set via VM-MIL. We quantize the SIFT features of each image class into visual words via K-Means, and the number of visual words in each class N2 is increased gradually from 200 to 1000, correspondingly the total number of visual words increases from 3000 to 15000. The result is in Table 3. It is shown that when the total length of visual words is 15000 we have the highest accuracy. Considering the calculation efficiency, we fix N2 to 1000 in the following experiments, thus the length of visual word set is set to 1500.

Table 3 The average classification accuracies of different lengths of visual word set

Length	3000	6000	9000	12000	15000
Accuracy	0.514	0.6098	0.6899	0.7401	0.8156

On the basis of visual word set with the length of 15000, we construct a visual phrase vocabulary whose length is about 25,000. The comparisons among the best results of VW-MIL, MVWL, MVPL and SPM are shown in Table 4.

Table 4 The average accuracies of different methods

SPM	81.3%
VW-MIL	81.5 %
MVWL	86.1%
MVPL	88.3%

Next we concentrate on the complexities of SPM and MVPL. For the experiment on Caltech 101, SPM follows the standard process as shown in literature (Lazebnik et al, 2006), i.e. an image is described by a 3-layer pyramid. Suppose the size of visual word group is 5600, then the dimension of image features is 5600*21=117600. Meanwhile HIK (Histogram Intersection Kernel) is used to measure the similarity of different images, whose calculation complexity is much higher than linear kernel calculation. In contrast, we select around 40,000 visual phrases via FP-Tree in MVPL and perform learning via L1-norm SVM. Thus the operation complexity of MVPL is lower than SPM.

CONCLUSIONS

In this article, we proposed a visual phrase generation method based on MIL framework and put forward MVPL based on this. Visual phrases are constructed from visual word sets in the same region and are capable to show the spatial relations of visual words. Thus visual phrases are more discriminative and descriptive. MVPL uses visual phrase vocabulary as the feature space to make the classification result more accurate. Experiments on two image sets have proved that our method is more effective than present classical algorithms.

.

REFERENCES

[1] Iliadis, M. ; Seunghwan Yoo ; Xin Xin ; Katsaggelos, A.K,Virtual touring: A Content Based Image Retrieval application , 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp.1 - 4,2013..

[2] Chucai Yi ; YingLi Tian,Localizing Text in Scene Images by Boundary Clustering, Stroke Segmentation, and String Fragment Classification , IEEE Transactions on Image Processing,Volume: 21 , Issue: 9,pp.4256 - 4268,2012.

[3] Maron O, Lozano-Perez T. A Framework for Multiple-Instance Learning [C]. Proceedings of Neural Information Processing Systems, 10: 570-576, 1998.

[4] Aissam Bekkari, et al., SVM Classification of Urban High-Resolution Imagery Using Composite Kernels and Contour Information, International Journal of Advanced Computer Science and Applications , vol. 4, no. 7, 2013.

[5] Wu Z, Ke QF, Sun J. 2009. Bundling features for large-scale partial-duplicate web image search[C]. In Proc. CVPR, 25-32.

[6] Hong Pan ; Yaping Zhu ; Qin, A.K. ; Liangzheng Xia,Mining heterogeneous class-specific codebook for categorical object detection and classification ,Image Processing (ICIP), 2013 20th IEEE International Conference on,pp. 3132 - 3136,2013..

[7] Liu D, Hua G, Viola P, Chen T. 2008. Integrated Feature Selection and Higher-order Spatial Feature Extraction for Object Categorization[C]. Proceeding of the 26th IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK.

[8] M.Iwahara, S.C.Mukhopadhyay, S.Yamada and F.P.Dawson, "Development of Passive Fault Current Limiter in Parallel Biasing Mode", IEEE Transactions on Magnetics, Vol. 35, No. 5, pp 3523-3525, September 1999.

[9] Zheng YT, Zhao M, Neo SY, Chua TS, Tian Q. 2008. Visual Synset: towards a Higher-level Visual Representation[C]. In Proc.CVPR, Achorage, Alaska, U.S.

[10] Yuan YS, Wu Y, Yang M. 2007. Discovery of Collocation Patterns: from Visual Words to Visual Phrases[C]. Proc. of the 25th IEEE Conference on Computer Vision and Pattern Recognition, 1-8.

[11] Li FF, Perona P. 2005. A Bayesian Hierarchical Model for Learning Natural Scene Categories [C]. Proceeding of the 23rd IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 524-531.

[12] Lazebnik S, Schmid C, Ponce J. 2006.Beyond Bags of Features: Spatial Pyramid Matching forRecognition Natural Scene Categories [C]. Proceeding of the 24th IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 2: 2169 – 2178。

[13] Lowe D G. 2004. Distinctive Image Features form Scale-invariant Keypoints [J]. International Journal of Computer Vision, 60(2): 91 – 110.

[14] G. Sen Gupta, S.C. Mukhopadhyay, Michael Sutherland and Serge Demidenko, Wireless Sensor Network for Selective Activity Monitoring in a home for the Elderly, Proceedings of 2007 IEEE IMTC conference, Warsaw, Poland, (6 pages).

[15] Zhang Q, Goldman S A. 2001. EM-DD: an improved multiple-instance learning technique. Advances in Neural Information Processing Systems, Cambridge, CA: MIT Press, 1073-1080.

[16] Huang X, Chen SC, Shy M, et. al. 2002. User concept pattern discovery using relevance feedback and multiple-instance learning for content-based image retrieval [C].MDM/KDD 2002 Workshop Edmonton, 100-108.

[17] Fergus R, Li Feifei, Perona P, Zisserman A. 2005. Learning Object Categories from Google'sImage Search [C]. Proceeding of the 10th International Conference on Computer Vision (ICCV), 1816 - 1823.

[18] Blei D, Ng A, Jordan M. 2003. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research. 3: 993-1022.

[19] Subhas Chandra Mukhopadhyay, and Chien-Hung Liu, "Designing an Integrated Curriculum Platform for Engineering Education: A Hybrid Magnetic Bearing System", International Journal on Technology and Engineering Education, Vol.7, No.1, pp. 17-31. July 2010.

[20] Cao LL, Li FF. 2007. Spatially Coherent Latent Topic Model for Concurrent Object Segmentation and Classification[C]. Proceeding of the 11th IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil, 1080 – 1087.

[21] Dalal N, Triggs B. 2005. Histograms of oriented gradients for human detection Computer [C]. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1(1): 886-893.

[22] S. C. Mukhopadhyay, G. Sen Gupta and S. Demidenko, "Intelligent Method of Teaching Eletectromagnetics Theory" Measurement Under Virtual Environment", International Journal on Smart Sensing and Intelligent Systems, Vol. 1, No. 2, June 2008, pp. 443-458.

[23] Kadir T, Brady M. 2001. Scale, Saliency and Image Description [J]. International Journal of Computer Vision, 45(2): 83-105.

[24] N. K. Suryadevara, S. C. Mukhopadhyay. R.K. Rayudu and Y. M. Huang, Sensor Data Fusion to determine Wellness of an Elderly in Intelligent Home Monitoring Environment, Proceedings of IEEE I2MTC 2012 conference, IEEE Catalog number CFP12MT-CDR, ISBN 978-1-4577-1771-0, May 13-16, 2012, Graz, Austria, pp. 947-952.

[25] Yanmin LUO, Peizhong LIU and Minghong LIAO, AN ARTIFICIAL IMMUNE NETWORK CLUSTERING ALGORITHM FOR MANGROVES REMOTE SENSING, International Journal on Smart Sensing and Intelligent Systems, VOL. 7, NO. 1, pp. 116 – 134, 2014

[26] Daode Zhang et al., RESEARCH ON CHIPS’ DEFECT EXTRACTION BASED ON IMAGE-MATCHING, International Journal on Smart Sensing and Intelligent Systems, VOL. 7, NO. 1, pp.321 – 336, 2014.

[27] Sean Dieter Tebje Kelly, Nagender Kumar Suryadevara, and S. C. Mukhopadhyay, "Towards the Implementation of IoT for Environmental Condition Monitoring in Homes" IEEE SENSORS JOURNAL, VOL. 13, NO. 10, OCTOBER 2013, pp. 3846-3853.