

DisIClass: Discriminative Frequent Pattern-Based Image Classification*

Sangkyum Kim, Xin Jin, Jiawei Han
University of Illinois at Urbana-Champaign
{kim71, xinjin3, hanj}@illinois.edu

ABSTRACT

Owing to the rapid mounting of massive image data, image classification has attracted lots of research efforts. Several diverse research disciplines have been confluent on this important theme, looking for more powerful solutions. In this paper, we propose a novel image representation method *B2S* (**B**ag to **S**et) that keeps all frequency information and is more discriminative than traditional histogram based bag representation. Based on *B2S*, we construct two different image classification approaches. First, we apply *B2S* to a *state-of-the-art* image classification algorithm *SPM* in computer vision. Second, we design a framework *DisIClass* (**D**iscriminative **F**requent **P**attern-**B**ased **I**mage **C**lassification) to utilize data mining algorithms to classify images, which was hardly done before due to the intrinsic differences between the data of computer vision and data mining fields. *DisIClass* adapts the locality property of image data, and apply sequential covering method to induce the most discriminative feature sets from a closed frequent item set mining method. Our experiments with real image data

show the high accuracy and good scalability of both approaches.

Categories and Subject Descriptors

I.4.10 [Image Processing and Computer Vision]: Image Representation; H.2.8 [Database Applications]: Data Mining, Image Databases

General Terms

Algorithms

Keywords

Image Classification, Image Mining, Image Categorization, Discriminative Pattern

1. INTRODUCTION

Image classification, or image categorization, has been intensively studied in several fields such as image processing, computer vision, pattern recognition, machine learning, and data mining. Some specialized image representations and classification models have shown great success in tasks such as face recognition, iris recognition and fingerprint recognition. However, these methods are designed for specific objects and usually require high quality and specialized training and testing images. Thus, they are not suitable for general image classification problems.

In the early days, global features like histograms of color, texture and edge information were used to express and classify images [2, 11, 22, 23]. The major drawback of this approach was that it captured not only the interesting target but also the noisy background. To avoid this problem, there were several works on subdividing an image into smaller blocks to exploit locality properties of image data [4, 18, 24]. They still got a problem that the schemes using fixed sized blocks were not robust on location translation which is common in real image classification applications.

Recently, there has been a trend to merge two or three fields together to take full advantage of the strength of each field. In computer vision, it

*This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (award number OCI 07-25070) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign, its National Center for Supercomputing Applications, IBM, and the Great Lakes Consortium for Petascale Computation. This work was also sponsored in part by the National Science Foundation (under grants IIS-09-05215, CCF-0905014, and CNS-0931975) and a KO-DAK gift scholarship. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MDMKDD'10, July 25th, 2010, Washington, DC, USA.

Copyright 2010 ACM 978-1-4503-0220-3 ...\$10.00.

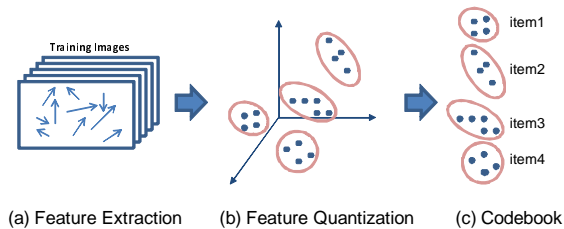


Figure 1: Codebook generation

has been tried to use artificial intelligence, statistics, and data mining techniques to gain higher performances. ‘*Bag-of-Words*’, one of those interdisciplinary paradigms, emerged in computer vision [17], which enabled the representation of an image as a bag of codewords and applied several techniques from other fields. The biggest difference of this paradigm from previous studies is the use of local features instead of global ones. Local features are computed from local interest regions, whose locations are determined by scale and affine invariant detectors.

In *Bag-of-Words* concept, an image is represented as a bag of codewords from a dictionary named *codebook*. To construct a codebook, we generally obey the following three steps:

First, feature detection is performed through all images. There are various methods on feature detection, including a regular grid method [17] and an interest point detection method [19]. The former is well known to have good results in natural scene categorization, but it cannot capture the salient features in an image which would be more useful in object recognition and image clustering.

Second, a feature is represented as a high-dimensional vector which becomes a descriptor of the feature. *SIFT* (Scale-invariant feature transform) [19] was introduced to detect and describe local image features based on the appearance of an object at particular interest points, which had very useful properties such as invariance to image scale and rotation. *SIFT* achieved the best performance in comparison with other descriptors [20].

Third, a vector quantization method is performed on the detected feature descriptors to generate the codebook. Each quantized vector becomes a codeword, and the codebook is composed of those codewords.

In Fig. 1, we illustrate the procedure to generate a codebook, which was used in this paper. We used interest point detection method and SIFT representation to extract features, and *k-means* clustering algorithm for vector quantization. Finding a better image processing technique such as designing a better feature extraction method or a better vector quantization method is beyond the scope of this paper.

Bag-of-Words is not a new concept in itself. It

was originated from natural language processing and information retrieval where a document was represented as an unordered collection of words, ignoring grammar and the order between words [16]. Applying this concept directly into different areas like computer vision contains several inherent problems.

First, the meaning of the number of replicates in an image is different from that of the term frequency in a document. A document is represented as a histogram of the terms used in it, and previous works in computer vision also have used the same framework: An image is represented as a histogram of the codewords. In a document, frequency measures the importance of the term, where a higher frequency implies more importance. In an image, we can barely find such relationships. In this paper, we propose a new representation *B2S* of an image and show how this representation leads to a higher classification accuracy.

Second, we cannot simply apply the same techniques we used for document classification to image classification. Unlike document data that has a sparse distribution of words, image data has a dense distribution of features. Due to this heavy density, it is hard or sometimes impossible to directly use data mining algorithms such as frequent itemset mining even for a small number of images. In this paper, we introduce a discriminative frequent item bag mining algorithm *DDB* which finds discriminative features without generating all frequent item bags.

Third, in an image, spatial information like absolute or relative locations are quite important to describe an object or a pattern in an image, while spatial information of terms are not considered in a document. Even worse, it is not robust on background noises if we consider the whole image as one document and ignore spatial information. To enhance the accuracy of the algorithms, we need to utilize spatial information in an image. There has been a few attempts to accomplish this goal [14, 21], but still there is much room left to be enhanced. In this paper, we utilize the locality property of image data and top-*k* probability voting to improve the performance.

In summary, the contributions of this paper are as follows.

- We analyze previous *Bag-of-Words* image representations, and develop a novel image representation method *B2S*.
- We show how *B2S* can be applied to existing image classification algorithms.
- We propose an image classification framework *DisIClass* based on a data mining approach by use of *B2S* representation. *DisIClass* utilizes the locality property of image data with the help of top-*k* probability voting scheme, and

solves all three problems of applying *Bag-of-Words* into computer vision.

- We perform experiments on two *B2S* based algorithms with real image data and compare with the *state-of-the-art* image classification algorithm *SPM* [9] to show that *B2S* representation really helps to achieve higher accuracy.

The rest of the paper is organized as follows. Section 2 presents a brief overview of related works. In Section 3, we present a formal definition and important properties of *B2S* representation, and show how to apply it to *SPM*. Section 4 proposes a frequent pattern mining based image classification framework *DisIClass* with the help of *B2S*. In Section 5, we report our experimental results which reveal high performances of our new approaches. Finally, we conclude the paper in Section 6 which summarizes our work on *B2S* and its two applications.

2. RELATED WORKS

In this section, we briefly introduce related works in computer vision and data mining.

2.1 Image Classification

The simplest way to describe an image is to use histograms of color, texture and edge information as global features. Many papers have been published in this direction. [22] studied binary image classification via color histograms using *k-NN* classifier. [23] proposed using histograms of colors and edge directions together with *Bayesian* classifiers for multiple class classification. [2] used color histogram features with *SVM* classifier for object classification. [11] used color information to construct a classification tree. The major drawback of those methods is that a global histogram representation could not discriminate between the interesting target and the background noises.

To avoid the problems of global histogram representations, subimage-based methods have been proposed which divided an image into several rectangular blocks to exploit local and spatial properties [4]. [24] partitioned an image into smaller blocks and represented them by use of wavelet coefficients in high frequency bands, and set a threshold to control a class prediction of each block. [18] proposed the *ALIP* system which used two-dimensional multiresolution *HMM* trained on color and texture features of image blocks. The problem of these methods is that fixed blocks are sensitive to the location change of objects in the images, thus they are not robust to location translation, which commonly happens in real images.

Image clustering and natural scene categorization fields have flourished due to the discovery of *SIFT* (Scale-Invariant Feature Transform) [19] feature, which was originated from the research of object recogni-

tion which required a good local feature to describe an object.

[17] introduced the whole framework of constructing codewords and a codebook from extracted features. It tested and compared several well-known feature extraction and description methods on a natural scene categorization problem.

There was a different approach that expressed an image as a graph [21]. A complete graph was constructed based on extracted features and their relationships. A well-known graph mining algorithm *gSpan* [25] was applied to the graph to mine interesting subgraphs. This approach also tried to utilize spatial information by representing an image as a graph with the help of data mining techniques, but it revealed a problem of high time and space complexity.

A new approach *SPM* [14] came out that utilized a kernel-based recognition method. It adapted the pyramid matching scheme [9] which used the spatial information indirectly. It could significantly improve performance over the basic *Bag-of-Words* image representation and even over methods based on detailed geometric correspondence. The basic idea of *SPM* is to repeatedly subdivide the image and computes histograms of local features at increasingly fine resolutions. Rough geometric correspondence is computed on a global scale by utilizing an efficient approximation method adapted from the pyramid matching scheme of [8]. Experiments on several benchmark data sets have shown that *SPM* achieves the *state-of-the-art* accuracy performance [7, 10, 14]. We perform the comparison experiments with *SPM* in Section 5 and show that our two approaches outperforms it.

2.2 Frequent Itemset for Image Mining

In the field of data mining, there has been several approaches to mine image data. They applied several data mining techniques on the assumption that a well-preprocessed image data is given as an input.

In [27], an image was represented as a set of recurrent items. A progressive deepening method was performed to mine association rules of predefined spatial relationships between objects such as *next-to* and *overlap*. Similar works appeared in spatial data mining area like spatial collocation mining which tried to find co-existing patterns of non-spatial features in a spatial neighborhood [12, 28].

There were several works that used a grid-based approach to explicitly code the location of objects in each image, which could generate spatial association rules by use of frequent pattern mining algorithm [3, 15]. Recently, there was a work to find semantically meaningful visual patterns based on a frequent pattern mining approach [26]. For each image, a *k-NN* clustering was performed on the code-

words to form a neighborhood as a meaningful visual patterns.

3. B2S REPRESENTATION

3.1 Foundation of B2S

Previously, all image mining algorithms based on *Bag-of-Words* paradigm used feature histograms to express an image. In other words, an image v could be represented as a vector of features x_1, x_2, \dots, x_n . Each feature x_i became a dimension, and its frequency (or the number of occurrences) v_i became the value of the corresponding feature dimension.

DEFINITION 3.1. *Suppose $X = \{x_1, x_2, \dots, x_n\}$ is a set of features in an n -dimensional vector space V_n of nonnegative integers. For any vector v in V_n , we denote $v = (v_1, v_2, \dots, v_n)$ where v_i is the frequency of x_i in v .*

Now, we propose a new discriminative representation $B2S$ of an image. The main idea is to transform a histogram based bag representation into a set representation. For each value of a feature in the bag representation, $B2S$ creates a new feature that corresponds to it.

DEFINITION 3.2. *Define c_i to be the cardinality of a feature x_i for $i = 1, \dots, n$. We define $B2S$ to be a function from V_n to V_M ($M = \sum_{k=1}^n c_k$) by $B2S(v) = (v_{11}, v_{12}, \dots, v_{1c_1}, v_{21}, \dots, v_{nc_n})$ for $v = (v_1, v_2, \dots, v_n) \in V_n$ where $v_{ij} = 1$ if $j \leq v_i$ and 0 otherwise.*

By cardinality of x_i , we mean the maximum frequency v_i of x_i . We illustrate the definitions mentioned above by the following example.

EXAMPLE 1. *Suppose x_1 and x_2 are two features of V_2 , and their cardinalities are 3 and 2 respectively. Then, we get five new $B2S$ features $x_{11}, x_{12}, x_{13}, x_{21}$ and x_{22} in V_5 . A vector $v = (2, 1)$ in V_2 becomes $(1, 1, 0, 1, 0)$ in $B2S$ representation.*

Note that $B2S$ representation defined in Def 3.2 is a one to one but not onto mapping by definition. Any v in the original space V_n can be mapped into a vector in V_M , but the contrary is not always true. In Example 1, we cannot map any inverse value of a vector $(0, 0, 1, 0, 0)$.

$B2S$ representation is useful since it is more discriminative than histogram based bag representation, while still keeping all original information, in the sense that any $B2S(v)$ can always be recovered to its original representation v because $B2S$ is a one-to-one function. We show several properties of $B2S$ in the following proposition and theorem.

PROPOSITION 1. *Let $v = (v_1, v_2, \dots, v_n)$ be a vector in V_n for a given attribute set $X = \{x_1, x_2, \dots, x_n\}$. Then, $B2S(v) = (v_{11}, \dots, v_{nc_n})$ preserves the frequency value v_i of each original attribute x_i .*

PROOF. We can recover v_i using $B2S(v_i)$ by

$$v_i = \sum_{j=1}^{v_i} v_{ij} = \sum_{j=1}^{c_i} v_{ij}$$

since $v_{ij} = 1$ for $j \leq v_i$ by Def 3.2. \square

Proposition 1 shows that $B2S$ representation keeps the frequency information of the original expression, which is the most important feature in histogram based representation. We also show that there is a property that makes $B2S$ different from previous bag representation in the following theorem and example. Before that, we first define separability of two data sets.

DEFINITION 3.3. *Let A and B be subsets of V_n . We define A and B to be **separable** if there exists a linear function f from V_n to \mathbb{R}^1 where $f(v) \geq 0$ for $v \in A$ and $f(v) < 0$ for $v \in B$.*

EXAMPLE 2. *Suppose $A = \{1, 3\}$ and $B = \{2\}$ in \mathbb{R}^1 . Then, there is no linear function from \mathbb{R}^1 to \mathbb{R}^1 that makes A and B separable. But, their $B2S$ representations $B2S(A) = \{(1, 0, 0), (1, 1, 1)\}$ and $B2S(B) = \{(1, 1, 0)\}$ become separable in \mathbb{R}^3 .*

THEOREM 1. *If two sets A and B are separable in V_n , then their $B2S$ representations $B2S(A)$ and $B2S(B)$ also becomes separable in V_M .*

PROOF. Since A and B are separable, there exists a linear function $f(x_1, x_2, \dots, x_n) = a_1x_1 + a_2x_2 + \dots + a_nx_n$ from V_n to \mathbb{R}^1 where $f(v) \geq 0$ for $v \in A$ and $f(v) < 0$ for $v \in B$ by definition. Let's define another linear function \tilde{f} from V_M to \mathbb{R}^1 by $\tilde{f}(x_{11}, x_{12}, \dots, x_{1c_1}, x_{21}, \dots, x_{nc_n}) = a_1(x_{11} + \dots + x_{1c_1}) + \dots + a_n(x_{n1} + \dots + x_{nc_n})$. Then, $\tilde{f}(\tilde{v}) \geq 0$ for $\tilde{v} \in B2S(A)$ and $\tilde{f}(\tilde{v}) < 0$ for $\tilde{v} \in B2S(B)$. \square

COROLLARY 1. *$B2S$ enables the data sets to be more separable which makes it more discriminative than previous histogram based bag representation, while containing all original information at the same time.*

PROOF. By Proposition 1, Example 2, and Theorem 1. \square

Note that the frequency of any feature in $B2S$ representation is either 0 or 1, which makes it a set expression. We can interpret $B2S$ as a conversion from **Bag to Set** representation.

In Fig. 2, we show the $B2S$ procedure. Let V_{trn} denote a set of images for training, V_{tst} a set of images for testing, and V the set of all images in a give data set. We use an integer to represent an item (or an attribute in computer vision term) in V . Suppose we have N distinct items $0, 1, \dots, N-1$ in V . We define a table of features *feature_table* at line 1 of Fig. 2 which stores a list of item expressions for

Procedure *B2S*

input1: preprocessed transactions in V_{trn} , and V_{tst}

input2: the number of distinctive features N in V

output: transactions in V_{trn} and V_{tst} in *B2S* representation

```

begin
1.  $feature\_table = \emptyset$ 
2.  $new\_feature = N$ 
3. for each transaction  $t \in V_{trn}$  {
4.   sort items in  $t$ 
5.   for each item  $i \in t$  {
6.     if  $feature\_table$  contains key  $i$  {
7.        $i\_count =$  current frequency of  $i$  in sorted  $t$ 
8.       replace  $i$  with  $feature\_table.get(i).get(i\_count)$ 
9.     } else {
10.      if  $i\_count$  is 1 {
11.        add  $(i, 1, i)$ 
12.      } else {
13.        add  $(i, 1, N)$ 
14.         $N = N + 1$ 
15.      }
16.      replace  $i$  with  $feature\_table.get(i).get(i\_count)$ 
17.    }
18.  }
19. }
20. Do the same procedure for transactions in  $V_{tst}$ 
    with  $feature\_table$  and  $new\_feature$  induced from above
end

```

Figure 2: *B2S* Procedure

each corresponding item. For example, if we have a transaction t of $\langle a, a \rangle$, then we will create a new item a' for the second item a of t , and t becomes $\langle a, a' \rangle$. Also, $feature_table$ will contain a list $[a, a']$ for the key a . To find out which item name is used for the duplicated items, we can simply reference this $feature_table$. At line 4, we sort the transaction to make the computation easier. At lines 3-19 of Fig. 2, we do the parsing and fill in the $feature_table$ with training image data V_{trn} . Once it is done, then we use this $feature_table$ to parse the test image data V_{tst} . The reason for this two-step approach is to simulate the real situations, since we would not be given the test data in advance.

Of course, *B2S* will increase the number of features in image data. To reduce the computational burden caused from this increment, we use *SVM* classifier [1] which works well even in a high dimensional vector space. In Section 5, we provide the experiment results of the comparison between bag and *B2S* representations, which reveals the discriminative power of our *B2S* representation.

Note that, *B2S* runs in a linear time of the number of total items in a data set, and scans the data set only once.

3.2 *B2S* Applications

B2S is quite useful in the sense that it can be easily applied to *state-of-the-art* algorithms in various domains including text mining and image classifica-

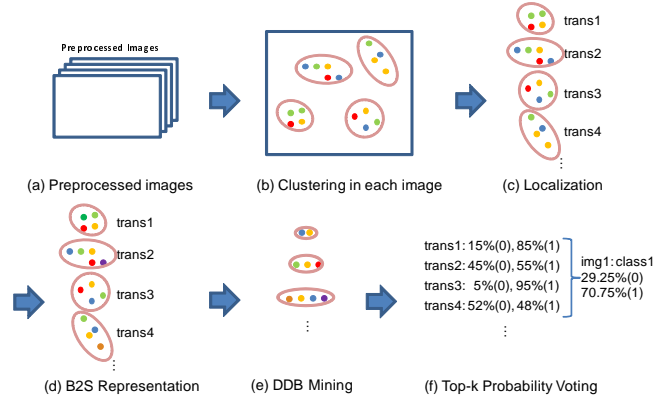


Figure 3: *DisIclass* Procedure

tion. It can be considered as a preprocessing part of the input data, which transforms it into the similar but more powerful format. In the experiment, we applied *B2S* to *SPM* [14] and show that *B2S* helps to improve its accuracy performance over 20%.

In Section 4, we describe a novel framework *DisIclass* that utilizes *B2S* as a connection between image classification in computer vision and frequent pattern mining in data mining domain. Lots of powerful data mining algorithms including frequent pattern mining could not be easily adapted into computer vision because of the intrinsic differences within their data. *B2S* enabled to translate the frequency information (used in computer vision and text mining fields) into a set representation (used in most data mining algorithms) without any loss of information.

4. DISICLASS FRAMEWORK

In this section, we propose a novel framework *DisIclass* for image classification. The whole procedure is described in Fig. 3. Once we get the preprocessed image data based on *Bag-of-Words* expression, we first perform the localization procedure for each image to utilize the locality property of image data. Then, we will have multiple transactions for each image data. We apply *B2S* method on those transactions to change the representation of images. Then, we use *DDB* algorithm to mine discriminative frequent item bags of the given training image data. Finally, we do the classification by use of *SVM* on those transactions, and conduct a top- k probability voting of transactions of each image to predict the class label of the test image data. The details of each step are described in the following subsections.

4.1 Localization

In image data, we observe the locality property that objects of an image are composed of a neighborhood of features. If we do not consider this prop-

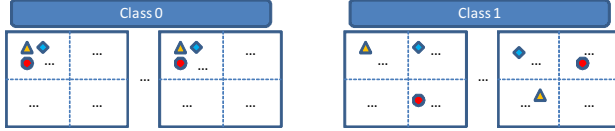


Figure 4: An Example of Locality Property

erty, we might miss to generate important discriminative patterns.

EXAMPLE 3. In Fig. 4, we see that a group D of a triangle feature, a diamond feature, and a circle feature forms a discriminative item set which will be used to judge whether an image is in class 0 or in class 1. But, in fact, all images contain D . Unless we pay attention to the locality property like ‘the item set D should be within a quarter of the image’, we will miss the most important pattern that discriminates class 0 from class 1.

There have been several approaches that utilize this locality property. In [14], a grid-based approach was proposed which shows a drawback in translated or rotated images. In [26], a k - NN clustering approach was proposed, where the clusters were overlapped and k had to be a small number (less than 10) to restrict the number of items in each transaction resulted in a low recall. In this paper, we use k -means algorithm to divide an image into several transactions, and change our image classification problem into a transaction classification problem. By use of k -means clustering algorithm, we are able to find discriminative patterns even for the translated or rotated images.

In fact, we can use any clustering algorithms for localization. The reason that there has been no work on these approaches for localization except k - NN (with a small k) is that sometimes these approaches might generate big-sized transactions which makes it impossible to mine all frequent patterns. In this paper, we mine the most discriminative feature bags without generating all frequent patterns, which works even with the big-sized transactions. The details are explained in Sec 4.2.

4.2 DDB Mining Algorithm

The study in [5] shows that discriminative power of some frequent patterns is much higher than that of single features, since a pattern which is a combination of several items is likely to capture more underlying semantics of the data. We cannot use frequency information to mine the most discriminative patterns, since their frequencies are neither too high nor too low.

To find them, we can simply generate all closed frequent patterns (if possible), and then choose discriminative patterns with a high information gain. But this approach does not work on image data since the features in image data are so dense that it

would generate exponential number of patterns that makes the computation infeasible. In [6], a new approach *DDPMine* was proposed to overcome this difficulty. Based on some pruning rules, it directly mined discriminative patterns by applying sequential covering method without generating all frequent item sets.

To mine a discriminative frequent item bag of image data, instead of mining item bags directly from the bag representations of image data, we first apply *B2S* to transform each image from a bag of items into a set of items, and then run *DDPMine*. We call this procedure *DDB* (Direct Discriminative Item Bag Mining). In this way, *B2S* enables not only to use a powerful data mining method but also to increase the accuracy of the image classification.

DDB mining showed a solution to the intrinsic problem of high density of image data. It enabled to overcome the restriction of the size of transactions, which was less than 10 [26], by more than an order of magnitude.

4.3 Classification with Top-k Probability Voting

In this paper, we used *SVM* classifier [1] since it has showed high quality results even for a data set in a high dimensional vector space. But, of course, any other high-quality classifier can be applied for the rest of the procedure.

Now, we are ready to classify images. All images are transformed into a set of transactions for localization in Section 4.1, changed into *B2S* representation and found the most discriminative item bags using *DDB* and integrated them into the image data in Section 4.2.

At first, it seems reasonable to perform the following two steps to use *SVM* classifier to classify images:

1. **Training step:** Train the transformed training image data set $(B2S + DDB)(V_{trn})$ and get its *SVM* model.
2. **Testing step:** Apply the trained model to predict the transformed testing image data set $(B2S + DDB)(V_{tst})$.

But, in fact, we have to classify images, not transactions. Moreover, many transactions contain noisy background information, which might lead to meaningless or misleading results. Therefore, we cannot merely use the predicted class labels of transactions of an image v and do a voting to determine class label of v , since there might be more transactions of noisy background information than transactions with meaningful patterns in v .

EXAMPLE 4. Suppose we want to do a binary classification of an image v in Fig. 5 which contains 3 transactions. Suppose that t_2 is the only transaction

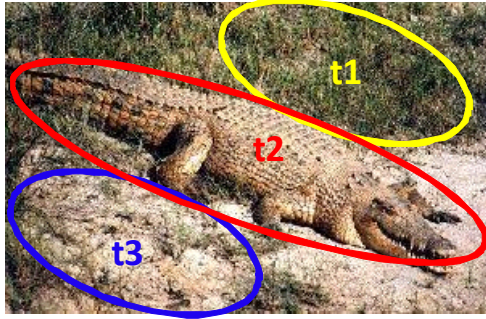


Figure 5: Example of Top- k Probability Voting

that is classified to class 0 while the others are classified to class 1. In this case, if we use the voting of the transactions, v will be misclassified to class 1. Therefore, we cannot simply use the voting of the transactions to determine the class label of v , because t_1 and t_3 do not contain the main discriminative features to determine the class of the whole image.

Due to the reasons mentioned above, we use the classifier not to predict the class label of each transaction but to predict the probabilities for a transaction to be in each class. In binary classification, we have observed that transactions containing garbage information predicted the probability for being in the correct class as around 50% whereas transactions containing meaningful information predicted the probability for being in the correct class with a high score around 90%. Now, we perform the top- k probability voting: the class label of an image is determined by the label which scored the highest average of top- k probabilities of each class label.

EXAMPLE 5. Suppose we want to do a binary classification of an image v in Fig. 5 which contains 3 transactions t_1 , t_2 and t_3 . Suppose that the probabilities of transactions to be in class 0 are $P_0(t_1) = 55\%$, $P_0(t_2) = 91\%$, and $P_0(t_3) = 52\%$. If we use the average value of top-2 probabilities of each class, then we get $P_0(v) = 73\%$ and $P_1(v) = 46.5\%$, and allocate v to class 0.

The reason that we use top- k instead of all values is to consider the case where the number of transactions per image is big. In that case, the accumulated effect of the background transactions might mislead classifier to a wrong result.

5. EXPERIMENTS

In this section, we describe our experiments to show the performance of two $B2S$ applications, $DisIClass$ and $SPM+B2S$. We conducted experiments with real image data sets to show the performances of our algorithms. Experiments were done on a PC

with a dual 3.4GHz Pentium D CPU and 1 GB RAM.

We used six image classes from Caltech-101 database [7], and conducted experiments on 3 pairs of classes for binary classification and compare our results with a *state-of-the-art* image classification algorithm SPM with 2-level pyramid kernel [14].

Based on the results in [14], we chose two classes (minaret, windsor chair) which achieved the best performances in SPM , two classes (ant, crocodile) with the worst performances, and we chose the other two classes (airplane, car) which contained many images. For data set A (minaret and windsor chair), we used 30 images from each class for training data and the remaining images (minaret:46, windsor chair: 26) for testing data. For data set B (ant and crocodile), we used 30 images from each class for training data and the remaining images (ant:12, crocodile:20) for testing data. For data set C (airplane and car), we used 50 images from each class for training data and 70 images from each class for testing data. We performed a binary classification for each data set. For the feature quantization, we used k -means clustering algorithm, and chose $k = 200$ based on the settings of [14]. The parameter setting of $DisIClass$ are mainly from [14] for the fair comparison with SPM since those settings would have been optimized to its performance.

In these experiments, we did not use any annotation information, and transformed all images into gray scale even when color images were available. We showed the accuracy of the prediction of test image labels for different $SIFT$ thresholds (0.015-0.03). If the threshold of $SIFT$ descriptor is set to be lower, then it detects and outputs a greater number of $SIFT$ features. Depending on the threshold, features are extracted in the number of hundreds or even thousands per image. For the following experiments, we define the accuracy of binary classification by

$$accuracy = \frac{num_cp}{num_tst_images} \quad (1)$$

where num_cp is the number of correct predictions for both classes and num_tst_images is the number of total test images of both classes.

5.1 Bag vs. B2S Representations

In this subsection, we show the usefulness of $B2S$ representation versus bag representation by comparing the accuracy performance of $SPM+B2S$ and SPM . As already shown in Corollary 1, $B2S$ representation has more discriminative power than bag representation, and the experiment results in Fig. 7 also reveal the same fact. For any data sets, $SPM+B2S$ shows higher accuracy performance than SPM itself, as expected. Even data set B which contains the most complicated images shows the accuracy improvement by around 10%.

Because of the page limit, we do not show all

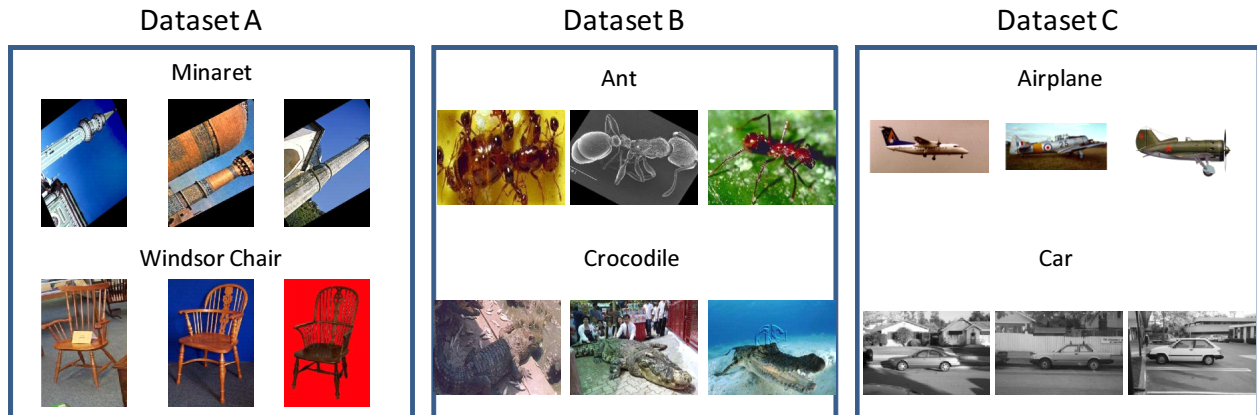


Figure 6: Misclassified or Bad Scored Images from *DisIClass*

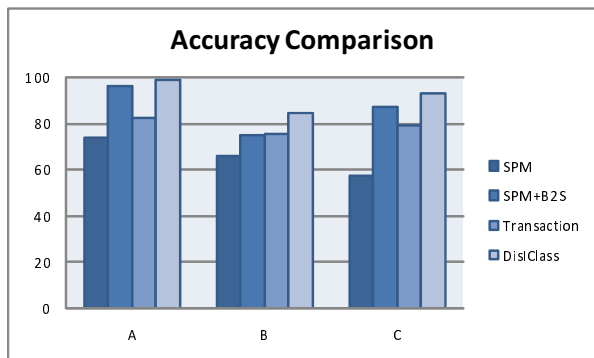


Figure 7: Accuracy Comparison of Four Different Algorithms (*SIFT* threshold 0.015)

experimental results in this paper, but we have to mention that we also observed that the difference of accuracies becomes bigger when the *SIFT* threshold becomes lower. This implies that the discriminative power of *B2S* representation becomes bigger when more features are extracted from an image. In general, it is better to extract more features not only for image classification but also for other algorithms on image data, and we believe it would also be useful to apply *B2S* representation to those tasks.

5.2 DisIClass vs. SPM and SPM+B2S

In this subsection, we compare the performances of *DisIClass* with *SPM* and *SPM+B2S*. We used 5 transactions for each image in localization procedure, and obtained top-4 probability voting for *DisIClass*.

Fig. 7 shows that *DisIClass* outperformed *SPM* and even *SPM+B2S*. In fact, *DisIClass* outperformed *SPM* in all cases, and outperformed *SPM+B2S* in all data sets for almost all thresholds. Considering that most images do not have any transformations, we can expect *DisIClass* will show bigger gaps compared to *SPM* and *SPM+B2S* in accuracy perfor-

mances on the data sets containing image transformations. *DisIClass* is a good example that applies frequent pattern mining techniques to image classification which was seldomly done before in computer vision, and it shows promising results.

DisIClass showed the lowest accuracy in data set *B* from three data sets. Since its images were quite complicated, it was hard to separate meaningful transactions from background transactions. Even in this hard data set, *DisIClass* showed better accuracy than the other two algorithms.

5.3 Transaction vs. Top-k Probability Voting (DisIClass)

In this experiment, we performed *k-means* clustering for each image with $k = 5$, and used top-4 probability voting to see how much the locality property of image data can improve the prediction accuracy. Since transactions are input data instead of images, we modified our accuracy function in Equation 1 for localization case by counting the number of transactions instead of the number of images. For the case of top- k probability voting, we used the accuracy function defined in Equation 1 because its results were based on an image unit.

As shown in Fig. 7, simply dividing image into several transactions was not enough to achieve higher accuracy. On the contrary, it frequently performed worse than not using the localization method, because many transactions of images were the noisy backgrounds. Merely predicting transactions based on those transactions really did harm the performance because it might confuse the classifier, and that is the reason we devised a novel way to predict correctly the class label of each image.

Noisy background transactions of each image achieved around 50% of prediction for both classes since their information gain were low. Based on this observation, we designed a top- k probability voting method for each image and it really outperformed the simple transaction-based approach in Fig. 7. It is be-

cause each image contained at least one discriminative transaction which determined the class label correctly, and for that transaction the information gain was high which resulted in a high probability (around 90%) of prediction. The probability voting method could really catch those discriminative subdivision and reflect them into correctly predicting image class label.

Considering the number of test images of data set B is small, we can mildly assume that the result of data set B shows that even if we use the probability voting scheme, the performance might be lower than the normal case if the image set contains a lot of noise inside.

5.4 Efficiency

In this subsection, we analyze the efficiency of two $B2S$ applications $SPM+B2S$ and $DisIClass$. Let m be the number of total items in the given data set, and n be the number of images.

Since $B2S$ creates new attributes, it seems that $B2S$ will make the whole process inefficient. But it is not quite true for image classification by the following two reasons: First, image data has a dense distribution of features and each feature has low frequency. Thus, $B2S$ creates a reasonable number of features which can be considered as $O(m)$. Second, there are classification algorithms that work efficiently even in a high dimensional space. In the experiments, we used SVM classifier which is claimed to run in linear time [13].

In theory, $SPM+B2S$ has the same time complexity with SPM , and it turned out to be true in all experiments which showed similar running times. So for the rest of the subsection, we focus on the efficiency of $DisIClass$.

Like other works in computer vision, the most time consuming part is the image preprocessing job including feature extraction and feature quantization. The other parts are negligible compared to it. Excluding preprocessing part, we found localization and DDB spend longer time than the rest of $DisIClass$ procedures. It takes $O(m)$ time for both $B2S$ and top- k probability voting procedures, which means that they are linear to the number of total items of the data set. For localization procedure, it takes $O(nP)$ times for localization procedure where P is the time taken for 2-dimensional clustering on an image, since each image needs to be localized. For a fixed $SIFT$ threshold, we could say that the number of items per image has an upper bound, and consider P as a constant. Thus, we can say that localization procedure is linear to the number of images. Even though DDB mining is not linear operation, we can still claim its efficiency by the experiment results of [6], and we could improve the efficiency even more by utilizing minimum support and maximum support thresholds. In this way, we are able to claim that $DisIClass$ is scalable.

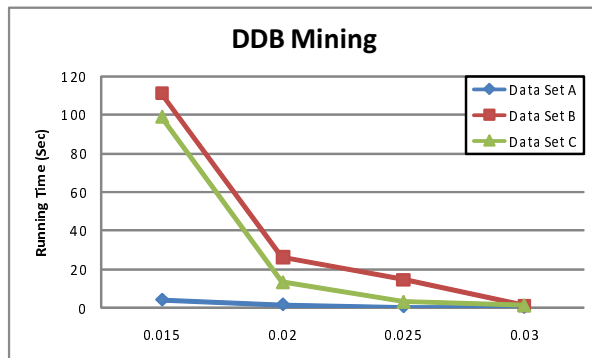


Figure 8: DDB Running Time

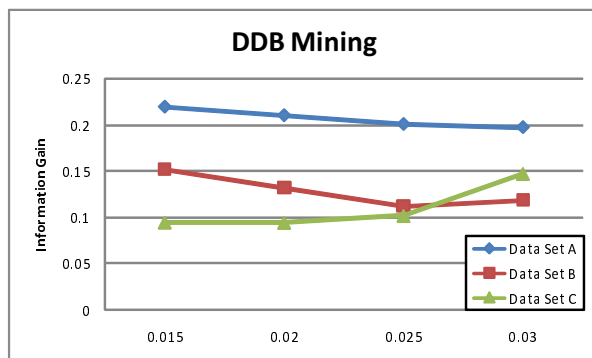


Figure 9: DDB Information Gain

We showed the DDB running time to get the first discriminative item bag in Fig. 8 and its information gain in Fig. 9, since the first pattern has the highest information gain, and it usually means the most critical one. As shown in Fig. 8, DDB mining works faster when the threshold becomes higher for both data sets. The reason is because the higher the threshold, the smaller the number of features are extracted for an image. Once the number of features becomes smaller, the number of items for each transaction also becomes smaller. For this reason, DDB mining can be done more efficiently.

One interesting observation is that the number of test images in data set A (72) is bigger than that in data set B (32). In general, a smaller data set should run faster than a larger one. So, in this case, DDB should mine data set B at least 2 times faster, but the result is rather different from what was expected. The reason is because data set B is harder to classify. Of course, we can directly check the images to see which data set is harder to classify. But, there are a couple of ways to prove it objectively. First, we can check the classification result and compare their accuracies. It was already shown previously in [14], and also we showed the result in Fig. 7. Second, we can see the difference of maximum information gain value of selected patterns between those two data sets. In fact, the discrimi-

native patterns from data set A got higher information gain than the patterns from data set B in Fig. 9. It means that mining the pattern with higher information gain enabled more pruning that resulted in a faster running time of DDB mining. In other words, if discriminative patterns from a data set get higher information gain, it would be easier to classify the data set with high accuracy and efficiency. Note that, the information gain scores are low in Fig. 9. It is because we used transaction count instead of image count, but they were enough to see the relative differences and trends.

6. CONCLUSION

In this study, we designed a new representation $B2S$ of image data to make it more discriminative than the previous histogram-based bag representation. Based on $B2S$, we proposed two image classification algorithms $SPM+B2S$ and $DisIClass$. The former is based on existing algorithm SPM and the latter is utilizing various data mining techniques. We showed three problems of the *Bag-of-Words* paradigm in previous computer vision studies, and solve all of them with this new framework $DisIClass$.

Our extensive experiments on different settings as well as our comparison of the results with a state-of-the-art image classification algorithm, SPM [9], have shown the high performances of our two $B2S$ applications.

We believe this new image representation has a high promise for effective image classification and object identification in many applications, and also gives an opportunity to use other powerful data mining techniques to mine image data.

7. REFERENCES

- [1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [2] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. In *IEEE Transactions on Neural Networks, special issue on Support Vectors*, 1999.
- [3] W.-T. Chen, Y.-L. Chen, and M.-S. Chen. Mining frequent spatial patterns in image databases. In *PAKDD '06: Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conference*, 2006.
- [4] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *J. Mach. Learn. Res.*, 5:913–939, 2004.
- [5] H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In *ICDE*, 2007.
- [6] H. Cheng, X. Yan, J. Han, and P. S. Yu. Direct discriminative pattern mining for effective classification. In *ICDE*, 2008.
- [7] L. Fei-Fei, R. Fergus, and P. Perona. Earning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR*, 2004.
- [8] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 2:1458–1465.
- [9] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, 2005.
- [10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.
- [11] J. Huang, S. R. Kumar, and R. Zabih. An automatic hierarchical image classification scheme. In *MULTIMEDIA '98: Proceedings of the sixth ACM international conference on Multimedia*, 1998.
- [12] Y. Huang, H. Xiong, S. Shekhar, and J. Pei. Mining confident co-location rules without a support threshold. In *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*, 2003.
- [13] T. Joachims. Training linear svms in linear time. In *KDD*, 2006.
- [14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
- [15] A. J. T. Lee, R.-W. Hong, W.-M. Ko, W.-K. Tsao, and H.-H. Lin. Mining spatial association rules in image databases. *Inf. Sci.*, 177(7):1593–1608, 2007.
- [16] D. D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, 1998.
- [17] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.
- [18] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1075–1088, 2003.
- [19] D. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV*, 1999.
- [20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.
- [21] S. Nowozin, K. Tsuda, T. Uno, T. Kudo, and G. H. Bakir. Weighted substructure mining for image analysis. In *CVPR*, 2007.
- [22] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *IEEE International Workshop on Content-based Access of Image and Video Databases*, 1998.
- [23] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, 2001.
- [24] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. In *VISUAL '00: Proceedings of the 4th International Conference on Advances in Visual Information Systems*, 2000.
- [25] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *ICDM*, 2002.
- [26] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *KDD*, 2007.
- [27] O. R. Zaiane, J. Han, and H. Zhu. Mining recurrent items in multimedia with progressive resolution refinement. In *ICDE*, 2000.
- [28] X. Zhang, N. Mamoulis, D. W. Cheung, and Y. Shou. Fast mining of spatial collocations. In *KDD*, 2004.