DISCRIMINATIVE SIGNATURES FOR IMAGE CLASSIFICATION

Ziming Zhang, Syin Chan, Liang-Tien Chia

Center for Multimedia and Network Technology, School of Computer Engineering
Nanyang Technological University, Singapore 639798
$\{zhan0154, asschan, asltchia\}@ntu.edu.sg$

ABSTRACT

Bag-of-words representation has shown to be a powerful technique for image classification. In this paper, we propose a new approach to discover the discriminability of each visual word (image feature) in the codebook for each image category. A general linear model (GLM) is employed to construct new histograms of the images which are the basis for image classification. We also discuss the relations between our approach and boosting approaches and non-negative matrix factorization (NMF).

Index Terms— discriminative, signature, image classification

1. INTRODUCTION

A popular technique in text classification, bag-of-words representation [1], has been applied to images in recent years. The basic idea of bag-of-words is to create a word-document co-occurrence matrix, where the visual words (image features) are the elements of the codebook formed by the training image data. These visual words can be considered independent of each other in the vector space. Our approach attempts to discover the *category discriminability* of each visual word in the codebook based on this matrix. Then these visual words are combined to form new visual terms with better discriminability which can result in the improvement of image classification. For combining visual words, we use a general linear model (GLM) to construct the visual term histogram of each image. After this, a classification method, *e.g.* SVM, Probabilistic Latent Semantic Analysis (pLSA) [2], is employed to classify the images based on these new histograms. In the following sections, we use "*VW*" and "*VT*" to refer to visual word and visual term respectively, and use "*signature*" to refer to the VW or VT histogram of an image.

The rest of the paper is organized as follows. In Section 2, our approach, *Discriminative Signatures (DS)*, is explained in details. In Section 3, we explain the relations between our approach and boosting approaches [3] and non-negative matrix factorization (NMF)[4]. In Section 4, we show our experimental results for image classification. Section 5 concludes the paper.

2. DISCRIMINATIVE SIGNATURES

In this section, we will explain our approach, *Discriminative Signatures (DS)*. The basic idea is to group different selected VWs to form more discriminative VTs so that the inter-category distance, S_B, can be maximized while the intra-category distance, S_W, can be minimized. Also we employ GLM to construct the new image signatures.

2.1. General linear model

A *general linear model (GLM)* is a model which explains the response variable y_i in terms of a linear combination of the explanatory variables $x_{i,j}$ plus an error term ϵ_i. That is,

$$y_i = \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + \cdots + \alpha_j x_{i,j} + \epsilon_i \tag{1}$$

where $\alpha_k (k \in \{1, \cdots, j\})$ is a coefficient of explanatory variable $x_{i,k}$.

In the form of matrices, a GLM can be described as below.

$$\mathbf{Y} = \alpha\mathbf{X} + \epsilon \tag{2}$$

where **Y** denotes the response variable matrix, **X** denotes the explanatory variable matrix, α denotes the coefficient matrix of **X** and ϵ denotes the error term matrix.

In our case, **Y** is the new *VT-document matrix*, **X** is the original *VW-document matrix*, and α is the *transform matrix*. We suppose $\epsilon = \mathbf{0}$, and our goal is to find an optimal α so that S_B can be maximized while S_W can be minimized as calculated using **Y**.

2.2. Objective function

We take the ratio of S_B to S_W as our main objective function, that is,

$$L = \frac{S_B}{S_W} \tag{3}$$

and α is optimal if it maximizes the ratio. So we have

$$L(\alpha) = \max_{\alpha}\{\frac{S_B}{S_W}\} \tag{4}$$

The definitions of S_B and S_W are

$$S_B = \sum_c (\mu_c - \overline{x})^T (\mu_c - \overline{x}) \tag{5}$$

$$S_W = \sum_c \sum_{r \in c} (x_r - \mu_c)^T (x_r - \mu_c) \tag{6}$$

where,

$$\mu_c = \frac{1}{N_c} \sum_{r \in c} x_r \tag{7}$$

$$\overline{x} = \frac{1}{N} \sum_r x_r = \frac{1}{N} \sum_c N_c \mu_c \tag{8}$$

Here, μ_c denotes the average signature of category c, $\overline{x}$ denotes the average signature of all the training images, x_r denotes the signature of image r in the word-document matrix $\mathbf{X}$, $r \in c$ means image r in category c, N_c is the number of training images in category c, and N is the number of all training images.

In our case, the matrix α should also satisfy the following conditions.

Condition 1: Non-negative Because the VTs are created by combining some of the VWs without subtraction, every element in α should be no less than zero.

Condition 2: Equal-sum The rows of α can be divided into C sets where C is the number of categories (one set, per category), and in each set the sum of the elements in each row should be approximately equal. This ensures that in each category each VT will have similar discriminability for this category.

So, our final objective function is

$$L(\alpha) = \max_{\alpha} \{ \frac{S_B}{S_W} \} \tag{9}$$

$$s.t. \left\{ \begin{array}{ll} \alpha_{ij} \geq 0 & i \in \{1, \cdots, m\}, j \in \{1, \cdots, n\} \\ \sum_{j=1}^{n} \alpha_{ij} = w_c & i \in \{1, \cdots, m_c\}, c \in \{1, \cdots, C\} \end{array} \right.$$

where m and n are the numbers of rows and columns of α, respectively, m_c is the number of rows in set c, and w_c is the pre-calculated sum of each row in set c (Please refer to Algorithm 1, Step 2 (a) and (b)).

2.3. Upper bound of $L(\alpha)$

Considering the complexity of our objective function, we would like to find a fixed upper bound of $L(\alpha)$ instead of its maximization. Let $p^c = \bigvee \{\mathbf{0}, \mu_c - \overline{x}\}$ and $q_r^c = \bigvee \{\mathbf{0}, x_r - \mu_c\}$ which are calculated from the training set of $\mathbf{X}$. Here, "$\bigvee$" denotes the operation that the maximal value of each element at the same position in the vectors is picked out. Then we can rewrite S_B and S_W, which are calculated from $\mathbf{Y}$, as below.

$$S_B^Y = \sum_c (p^c)^T \alpha^T \alpha p^c = \sum_c \sum_{i,j,k} p_i^c p_j^c \alpha_{ki} \alpha_{kj} \tag{10}$$

$$\begin{aligned} S_W^Y &= \sum_c \sum_{r \in c} (q_r^c)^T \alpha^T \alpha (q_r^c) \\ &= \sum_c \sum_{r \in c} \sum_{i,j,k} q_{ir}^c q_{jr}^c \alpha_{ki} \alpha_{kj} \end{aligned} \tag{11}$$

where p_i^c is the i^{th} element in signature p^c, q_{ir}^c is the i^{th} element in signature q_r^c, α_{ki} is the element in α whose row is k and column is i.

Now we will prove that a fixed upper bound of $L(\alpha)$ exists.

Statement 1 *There exists a fixed upper bound of $L(\alpha)$.*
Proof:
Let $\beta_c = \frac{\sum_{r \in c} \sum_{i,j,k} q_{ir}^c q_{jr}^c \alpha_{ki} \alpha_{kj}}{\sum_{r \in c} \sum_{i,j,k} q_{ir}^c q_{jr}^c}$, where β_c is a positive constant when every q_{ir}^c and α_{ki} are known. Then S_W^Y can be rewriten in the following way.

$$S_W^Y = \sum_c \beta_c \sum_{r \in c} \sum_{i,j,k} q_{ir}^c q_{jr}^c \tag{12}$$

Further, let $\widehat{q}_i^c = \sqrt{\frac{1}{N_c} \sum_{r \in c} (q_{ir}^c)^2}$ and $\lambda_c = \frac{\sum_{r \in c} \sum_{i,j,k} q_{ir}^c q_{jr}^c}{\sum_{r \in c} \sum_{i,j,k} \widehat{q}_{ir}^c \widehat{q}_{jr}^c}$, where λ_c is a positive constant when every q_{ir}^c is known. So Equ.(12) can be rewritten as below.

$$S_W^Y = \sum_c \lambda_c \beta_c \sum_{i,j,k} \widehat{q}_i^c \widehat{q}_j^c \tag{13}$$

When every $\widehat{q}_i^c$ is known, Equ.(13) can be rewritten as below.

$$S_W^Y = \frac{1}{\theta} \sum_{c,i,j,k} \widehat{q}_i^c \widehat{q}_j^c \tag{14}$$

where θ is a positive constant. Therefore,

$$\begin{aligned} L(\alpha) = \frac{S_B^Y}{S_W^Y} &= \frac{\sum_{c,i,j,k} p_i^c p_j^c \alpha_{ki} \alpha_{kj}}{\frac{1}{\theta} \sum_{c,i,j,k} \widehat{q}_i^c \widehat{q}_j^c} \\ &\leq \theta \sum_{c,i,j,k} \frac{p_i^c p_j^c \alpha_{ki} \alpha_{kj}}{\widehat{q}_i^c \widehat{q}_j^c} \\ &\leq \theta \sum_{c,i,j,k} (\frac{p_i^c \alpha_{kj}}{\widehat{q}_i^c})^2 \end{aligned} \tag{15}$$

The equation holds only when

$$\alpha_{ki} = \frac{p_i^c}{\widehat{q}_i^c} \tag{16}$$

Then we can get the following inequality.

$$L(\alpha) \leq \theta \sum_{c,i,j,k} (\frac{p_i^c p_j^c}{\widehat{q}_i^c \widehat{q}_j^c})^2 \tag{17}$$

This is a fixed upper bound of $L(\alpha)$.

Here we call $\alpha_{ki} = \frac{p_i^c}{\widehat{q}_i^c}$ the *category discriminability of VW_i.*

2.4. Transform matrix α

Here we will address this question: how to create the transform matrix, α, so that it can satisfy both conditions in (2.2). Let us recall a set of classical problems: *subset sum problems (SSP)*[5].

Definition 1 *Subset sum problems*
Given positive integers $c_1, \cdots, c_m, s$, we wish to solve the equation $\sum_{i=1}^{m} c_i x_i = s$ with $x_1, \cdots, x_m \in \{0, 1\}$.

Our problem is similar to SSP, replacing integers with rational numbers. As we know, SSP is *NP-Complete*, and the computational complexity is exponential in the smaller of two parameters, the number of decision variables, U, and the precision of the problem, V. In our case, U is the size of the original visual codebook which is usually quite large, and V should be as high as possible. So even if we use optimal algorithms for SSP to solve our problem, it still needs a lot of computational time.

Thus we propose an algorithm to create α efficiently. We use *greedy search strategy* to find the elements of α such that α satisfies both conditions. Also we need to pre-define a parameter, *Category Row Number (CRN)*, η, which limits the number of rows in α to $\eta * C$. The computational complexity is linear in U. The algorithm is described below.

Algorithm 1 *Greedy search strategy for computing α.*

```
1. Pre-define CRN, $\eta$, to control the number of rows in $\alpha$.
2. For category c,
    (a) Referring to Eq. (16), calculate $\frac{p_i^c}{\overline{q}_i^c}$ for each VW
        using the training VW-document matrix.
    (b) Calculate $w_c$ using $w_c = \frac{1}{\eta} \sum_i \frac{p_i^c}{\overline{q}_i^c}$.
    (c) Sort $\frac{p_i^c}{\overline{q}_i^c}$ in descending order, and then select one item
        at a time beginning from the top.
    (d) Beginning from the first row of $\alpha$, fill the selected
        element $i$ in column $i$ of the row. Fill the next row when
        the sum of the present row just exceeds $w_c$.
3. Repeat Step 2 for all the categories.
```

3. RELATION TO OTHER APPROACHES

Here we compare our approach with boosting approaches [3] and non-negative matrix factorization (NMF) [4].

Our method can be considered as a simplified version of boosting approaches. In traditional boosting approaches, an additive model is used to create response variables by combining explanatory variables. Some adaptive methods, such as Expectation Maximization (EM), are usually employed to learn the parameters in the additive model. However, in our approach we use statistical methods instead of adaptive methods to learn the parameters, which makes our approach much faster. Also in our approach we take the category discriminability of each VW as the weight, similar to those used in boosting approaches, and combine the boosted VWs together.

In NMF, it decomposes the word-document matrix into more basic part-document matrices, which are also linear representations of non-negative data. However, due to the different purposes, the learning methods are different. NMF uses adaptive methods while ours uses statistical methods. NMF tries to describe objects with more detail features, so it is more suitable for object recognition, such like face recognition. In contrast, our approach tries to describe objects with the combination of different discriminative VWs, so our approach is more suitable for image classification.

4. EXPERIMENTS

We use the Caltech image database [6] to evaluate our approach. It comprises five categories: motorbikes (826 images), faces (450 images), airplanes (1074 images), cars (rear) (1155 images) and background (900 images).

In our experiments, all images are gray-scale and 300-pixel wide. Each dataset is split randomly into two separate sets with equal size, training set and test set, for each run of the program. To describe each image, we extract the local image regions using the saliency region detector[7]. Each local region is resized to 16*16 pixels, and further divided into 4*4 smaller regions, each with 4*4 pixels. Each smaller region is then represented by a 8-dimension gradient bins similar to SIFT descriptor[8]. Concatenating these 16 smaller region descriptors together and normalizing the vector, we obtain a normalized 128-dimension vector for each local region. Then we use *k-means* to cluster the region descriptors in the training set to form a codebook and use VQ to create signatures of all the images based on the codebook. After using our approach to construct the new signatures, pLSA (code from [9])is employed to classify the images. For the EM in pLSA, the maximal number of iterations is 100, and the minimal allowable likelihood change is 1. We fix the size of the codebooks to 1000, and *CRN* is 15 empirically. All the results shown are the average of 50 runs.

We assess our approach in: (a) discriminability of signatures, (b) image classification.

4.1. Discriminability of signatures

Table 1 shows the extent of improvement in the discriminability of signatures, where *OS* is short for original signatures. The results shown are the values of L, calculated by Eq.(3), (5) and (6) using inter-category distance and intra-category distance, and the percentage of improvement (PI). The values in the *OS* column and *DS* column are calculated by the original signature matrix (X) and the discriminative signature matrix (Y), respectively. We can see that the discriminability of the signatures has been enhanced greatly through our approach.

Table 1. Comparison of discriminability

Category	Value of $L(\alpha)$ ($*10^{-4}$)		
	OS	DS	PI (%)
Motorbikes + Background	0.8	***5.6***	600
Faces + Background	1.4	***8.1***	479
Airplanes + Background	1.2	***6.5***	442
Cars (rear) + Background	1.0	***6.1***	510

4.2. Image classification

We compare our results with those in [6] using Receiver Operating Characteristic (ROC) curves.

For binary classification (*i.e.*foreground vs. background), the results of Areas Under the Curves (AUCs) are shown in Table 2. Comparing *DS* with *OS*, we can conclude that our approach (*DS*) does improve the performance of image classification greatly, and also our results are better than [6] except for the *airplanes* category. The reason is probably because relatively fewer image regions are extracted from the *airplanes* category than from other categories, which can affect the performance of classification greatly. On average, 37 regions are extracted from an image in the *airplanes* category while 46, 59, 66, 46 regions, respectively, for category *faces, motorbikes, cars(rear)* and *background*. Similar observations are made in multi-class classification.

For multi-class classification, we show and compare our results with those in [6] using different categories and different numbers of topics in pLSA. See Table 3, where "Cat." denotes the categories used in the experiments, "T" denotes the number of topics in pLSA, "M", "F", "A", "CR" and "bg" denote the categories of motorbikes, faces, airplanes, cars(rear) and background, "Ave. acc." denotes average accuracy. From the average accuracy, we can see that except for the first case, our results are all better than those in [6]. Moreover, with the increase in the number of topics under "4 + bg" categories, our results are more stable. This demonstrates that based on our discriminative signature matrix, pLSA can also find suitable topics for the categories when the number of topics are more than that of the categories.

Table 2. Comparison of binary classification (%)

Dataset	DS	OS	PI(%)	Sivic et al.[6]
Motorbikes	***96.8***	86.2	12.3	84.6
Faces	***96.5***	89.3	8.1	94.7
Airplanes	90.1	89.6	0.6	***96.6***
Cars (rear)	***99.1***	98.7	0.4	78.6

Table 3. Comparison of multi-class classification (%)

Cat.	T	*DS*					Ave. acc.	
		M	F	A	CR	bg	*DS*	[6]
4	4	98.6	99.4	86.8	98.9	×	96	***98***
4 + bg	5	94.5	97.8	86.1	98.2	70.4	***89***	78
4 + bg	6	95.1	98.0	86.4	93.5	79.0	***90***	76
4 + bg	7	96.8	98.1	86.9	97.1	69.2	***90***	83

5. CONCLUSION

In this paper, we propose a new approach, *Discriminative Signature*, to create more discriminative signatures for images by combining selected visual words for each category. Our results show that the discriminability of the visual terms is greatly enhanced, leading to better performance in image classification.

6. REFERENCES

[1] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual categorization with bags of keypoints," in *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[2] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Hingham, MA, USA, 2001, vol. 42, pp. 177–196, Kluwer Academic Publishers.

[3] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, June 2004, vol. 2, pp. 762–769.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[5] M. R. Garey and D. S. Johnson, "Computers and intractability: A guide to the theory of np-completeness," New York, NY, USA, 1990, W. H. Freeman & Co.

[6] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *ICCV*, 2005.

[7] T. Kadir and M. Brady, "Saliency, scale and image description," in *IJCV*, November 2001, vol. 45, pp. 83–105(23).

[8] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, 2003, vol. 20, pp. 91–110.

[9] R. Fergus, "Iccv short course 2005," in $http://people.csail.mit.edu/fergus/iccv2005/bagwords.html.$