# A NOVEL IMAGE CLASSIFIER BASED ON GAUSSIAN MIXTURE LANGUAGE MODEL

*Wei Wu, Guanglai Gao*

Department of Computer Science, Inner Mongolia University, Huhhot, 010021, China

**ABSTRACT**

In this paper, we propose a novel Gaussian Mixture Language Model to address the issues of the traditional bag of visual words (BoVW) based model. We firstly take full advantage of image semantic information to learn a new distance metric which can achieve the minimal loss of image information, and then we train Gaussian Mixture Models (GMM) using this distance metric. Given a test image, a visual document is firstly constructed using this codebook, and then its category is determined by estimating the maximum probability using the language model under a specific category. Experiments show that the codebook generated by our method can effectively reflect the image semantic information and highly suitable the language model, and confirm that the proposed method is satisfactory and competitive in comparison with the traditional BoVW based method as well as other state of the art methods.

***Index Terms***— Image classification, Bag of visual words, Gaussian mixture models, Distance metric learning, Language model

## 1. INTRODUCTION

Image classification is a problem of great interest in both research and applications. Although it has been studied for many years, it is still challenging within the field of multimedia and computer vision. The bag of visual words (BoVW) models [1-5], one of the most successful models for the object categorization and image classification tasks, has generally shown promising performance, and thus has been widely adopted in these fields. But the BoVW based methods are subject to some limitations. One is the ignorance of spatial information of the feature patterns. Recently, many advanced methods have been proposed to address this problem, and have achieved good performance: see, e.g., sparse coding [2], local coordinate coding (LCC) [3], super-vector coding [4], and etc.

Another limitation of BoVW is that much valuable information is missed when constructing the codebook (visual words) by clustering local features in the Euclidian space without considering the correlations among the visual words. Recently, most of the research work focused on enhancing the discrimination of the codebook to alleviate this issue. For instance, Jurie and Trigs [6] proposed a scalable radius based clustering method, Wu and Rehng [7] used histogram intersection kernel to create codebooks, Gemert et al. [8] studied soft-assignment codebook model, Zhou and Fan [5] presented a joint dictionary learning algorithm (JDL), and Jiang et al. [9] reported a label consistent K-SVD algorithm to learn a discriminative dictionary. But all of these methods used k-means based clustering algorithms to construct codebooks, which is difficult to avoid information loss, and didn't consider the spatial correlation between the different visual words. We think that if we could minimize the information loss when constructing the codebook, and meanwhile, take full advantage of the correlations among different visual words, the classification performance would be significantly improved.

In this paper, we adopt a probabilistic framework and a semantic distance to address the issues in face of BoVW. An image is represented by a set of local features, which can be interpreted as a probabilistic distribution in the feature space. Hence, we first use image local features with semantic information to learn a new distance measure through distance metric learning (DML), and construct codebooks by Gaussian Mixture Models (GMM) trained by this new distance measure, which can minimize the image semantic information loss. Then, we consider the spatial correlations among the visual words in an image, which form a document. Finally, for all these visual documents, we make use of language model (LM) widely used in the field of text information retrieval for image classification. We term this method as Gaussian Mixture Language Model. Compared with BoVW, our method takes into account both the information loss and the correlations among the visual words. Experiments confirm the effectiveness of our method.

There are some existing studies related to ours. Works [10-13] were based on probabilistic frameworks: [10] used GMM to build codebooks; [11] estimated a GMM for each image and used its parameters as the feature; [12] presented a global Gaussian features method; [13] used multinomial distribution to construct BoVW. All of these methods didn't leverage any image semantic information. Works [14-19] introduced semantic information for DML: [14-15] used a DML based k-means algorithm to build codebooks; [16-17] calculated image similarities directly through local features measured by DML; [18-19] adopted a set of different

features for DML. The difference between our method and the above DML based methods is that we combine a semantic based distance metric with GMM. Recently, the study of visual words spatial correlation for image classification is relatively few. [20] proposed a traditional LM for image classification, [21] presented a word spatial arrangement (WSA) strategy, but the experimental results were not satisfactory. In our method, we utilize a distance metric based GMM to construct visual documents, then apply it to LM for classification, and achieve satisfactory performance.

This paper is organized as follows. Section 2 describes the DML using semantic information, and section 3 introduces visual words generation based on GMM with a new distance metric, and then is the language model. Section 4 reports experimental results. Finally, we conclude our work and shed light on the future work in section 5.

## 2. SEMANTIC DISTANCE METRIC LEARNING

In order to preserve the semantic information and minimize the semantic loss in the codebook generation process, we introduce a novel distance metric learning scheme using image segmentation semantic information. The objective of DML (distance metric Learning) is to find an optimal Mahalanobis metric $A$ from training data with class labels or general pairwise constraints [19]. In our method, we extract the pairwise constraints from training images for distance metric learning. The pairwise constraints come from well segmented images. For example, the object region information of each image is provided by Caltech101 dataset [22]. We formalize the representation of the features pairwise constraints set $\{(x_{i1}, x_{i2}, y_i)\}_{i=1}^{N}$, where $x_{i1}$ and $x_{i2}$ are two $d$-dimensional features. And if both $x_{i1}$ and $x_{i2}$ are on the same semantic parts of objects, then $y_i = 1$, otherwise $y_i = -1$. It is worth noting that how to select pairwise constraints can greatly effect the classification performance. We comply with such selection criterion: the features $x_{i1}$ and $x_{i2}$ are of the same semantics but with large distance in Euclidean space, or vice versa.

Given the pairwise constraints information, the goal of our task is to learn a distance metric $A$ to effectively measure distance between any two visual features $x_{i1}$ and $x_{i2}$, following formula can represent this framework:

$$d(x_{i1}, x_{i2}) = \sqrt{(x_{i1} - x_{i2})^T A (x_{i1} - x_{i2})} \quad (1)$$

To find an optimal metric $A$, the distances between visual features of the same semantics should be minimized, and meanwhile distances between features of different semantics should be maximized. Based on this principle, we formulate this distance metric learning problem into the following optimization:

$$\begin{aligned} \min_{A,b} \quad & \sum_{i=1}^{N} y_i (\| x_{i1} - x_{i2} \|_A^2 - b) + \frac{\lambda}{2} tr(A^T A) \\ s.t. \quad & \sum_{i=1}^{N} y_i (\| x_{i1} - x_{i2} \|_A^2 - b) \le 1 \\ & A \ge 0, \quad \| A \| = 1/\sqrt{\lambda} \end{aligned} \quad (2)$$

where $\| \bullet \|_A$ is the Mahalanobis distance between two features under metric $A$. Parameter $\lambda$ is a constant, $b$ is a threshold. We use a stochastic gradient search algorithm to solve this optimization problem. The algorithm is described in following:

**Semantic Distance Metric Learning Algorithm**

Input: pairwise constraints $\{(x_{i1}, x_{i2}, y_i)\}_{i=1}^{N}$;
parameter $\lambda$, and learning rate parameter $\gamma$;
Procedure:
- Initialize $A = I$, $b = b_0$, iteration $t = 1$;
- **repeat**

**1** $\lambda = \lambda / t$, $t = t + 1$

**2** $C_t = \{(x_{i1}, x_{i2}, y_i) \mid y_i (\| x_{i1} - x_{i2} \|_A^2 - b) < 1\}$

**3** $f(A, C_t) = \sum_{(x_{i1}, x_{i2}, y_i) \in C_t} y_i (\| x_{i1} - x_{i2} \|_A^2 - b) + \frac{\lambda}{2} tr(A^T A)$

**4** Compute gradients: $\nabla_A f(A, C_t)$

**5** Compute gradients: $\nabla_b f(A, C_t)$

**6** update $A$ and $b$:
$A = A - \frac{\gamma}{t} \nabla_A f, b = b - \frac{\gamma}{t} \nabla_b f$

**7** constraint $A$ as positive semi-definite:
$A \leftarrow \sum_i \max(0, \lambda_i) \phi_i \phi_i^T$

**8** satisfy $\| A \| = 1/\sqrt{\lambda}$ : $A \leftarrow \frac{1}{\sqrt{\lambda}} \frac{A}{\| A \|}$

- **until** convergence

Output: metric $\boldsymbol{A}$, threshold $\boldsymbol{b}$.

In step 7, $\lambda_i$ and $\phi_i$ denote the $i$th eigenvalue and eigenvector of $A$. The algorithm is an iterative process, its computational complexity is determined by the product of the number of iteration and the size of training data, namely $O(tN)$. Empirically, this iterative algorithm converges quickly with no more than 5 iterations.

## 3. GAUSSIAN MIXTURE MODELS WITH DML

### 3.1. Visual words based on GMM

Most of the methods of generating the visual words have focused on the k-means clustering or its variations. In this paper, we utilize the GMM to construct the codebook. Each Gaussian component of GMM represents a visual word. Differetly from the traditional GMM [10-12], we adopt our semantic distance metric $A$ to build GMM. Moreover, for each image category, we respectively train a GMM, and

obtain a set of visual words. Finally, we form a global codebook by gathering all the visual words from all the image categories. Note that we share the distance metric $A$ among all the GMMs.

We model a GMM for the local feature $x$ of an image, and its form is:

$$p(x) = \sum_{k=1}^{M} w_k N(x \mid \mu_k, \Sigma_k), \quad (3)$$

where $x$ is a $d$-dimensional local feature, in our experiments, we use densely sampled SIFT features [1, 23], $w_k$, $\mu_k$ and $\Sigma_k$ denote the weight , mean vector and covariance matrix of the $k^{\text{th}}$ Gaussian component respectively, and $M$ is the total number of Gaussian components. Our model is different from the traditional GMM, in that we introduce a distance metric $A$ for the above Gaussian distribution:

$$N(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} \mid \Sigma_k \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2\Sigma_k}(x-\mu_k)^T A (x-\mu_k)\right) \quad (4)$$

This means that we add the semantic information to GMM, and make it represent image content better than general GMM. The GMM parameters are derived by using the Expectation Maximization (EM) algorithm. In order to reduce the number of parameters to be estimated so as to alleviate the computational cost for parameter estimation, we use diagonal covariance matrices $\{\Sigma_k\}_{k=1}^{M}$, which has been proved to be effective and computationally efficient [11].

When GMM is trained for each image category, we obtain $N$ GMMs, where $N$ is the number of image categories, and for simplicity, we share the same $M$ number of component for each GMM. Thus we totally obtain the $N \times M$ size of codebook for the whole image categories. Now, for a specific local feature $x$ of an image, we assign it to a visual word (a Gaussian component) which has maximum posterior probability for all $N \times M$ Gaussian components:

$$V(x) = \arg\max_{k,i} p_k(x, C_i), \quad k = 1, \cdots, M; i = 1, \cdots, N$$
$$p_k(x, C_i) = \frac{w_k N(x \mid \mu_k, \Sigma_k, C_i)}{\sum_{j=1}^{M} w_j N(x \mid \mu_j, \Sigma_j, C_i)} \quad (5)$$

where $V(x)$ denotes the visual words generating function, $p_k(x,C_i)$ is a posterior probability for the $k^{\text{th}}$ Gaussian component in GMM generated by the image category $C_i$. For the $N \times M$ size of codebook, we let: $v_1$, $v_2$, ... ,$v_{k,i}$=$v_{(i-1)\times M+k}$ , ... ,$v_{N\times M}$ , where $i$ and $k$ mean the $k^{\text{th}}$ Gaussian component in the $i^{\text{th}}$ GMM. So we assign the local feature $x$ to a visual word $V(x) = v_{(i-1)\times M+k}$. Thus, for each image, we can have a visual document constituted by a sequence of visual words using formula (5).

### 3.2. Visual language model

Traditional BoVW model does not consider the spatial correlation among the visual words, which might provide additional information to help image classification. So, we can use visual documents to construct language model.

When constructing visual documents, we only consider the visual word and its neighbor using bigram model, which is mainly from the viewpoint of computational cost and efficiency [24].

In our bigram model, each visual word is conditionally dependent on its one neighbor only. When all the bigram models corresponding to each image category are constructed, the new test image is then assigned to the most probable category by maximizing the posterior probability:

$$C^* = \arg\max_{C_k} p(C_k \mid D_{newImage}) = \arg\max_{C_k} \prod_{v_i v_j \in D_{newImage}} p(v_i \mid v_j, C_k) p(C_k) \quad (6)$$

where $D_{newImage}$ denotes the visual words document of the new test image, $p(C_k)$ is a prior probability for the $k$th category. $P(v_i \mid v_j, C_k)$ denotes the number of co-occurrence of $v_i$ and its left neighbor $v_j$ in category $C_k$.

## 4. EXPERIMENTS AND RESULTS

We use the Caltech101 dataset to carry out our experiments. The Caltech101 dataset [22] contains 101 classes, including animals, vehicles, flowers, etc., with high shape variabilities. Particularly, each image is provided with object segmentation information, which is the outline of each object in these images. So we can conveniently extract the semantic information (namely pairwise constraints set $\{(x_{i1}, x_{i2}, y_i)\}_{i=1}^{N}$) for the distance metric learning. We partition the whole dataset into 5, 10, 15, 20, 25, 30 training images per class and according, there are no more than 30 testing images per class. We repeat the experimental process by 5 times with different and randomly selected training and testing images to obtain reliable results. And the evaluation target we used is the average classification accuracy for all categories.

The 128-dimentational SIFT [1, 23] features extracted from 16×16 pixel patches were densely sampled from each image on a grid with the step size of 8 pixels. The experiments include two parts.

### 4.1. The performance effect of DML and codebook size

Firstly, we test the effect with or without the distance metric on GMM and k-means, and the performance of different sizes of visual codebook using our language model classifier. In experiments, we fix the number of training images in each class to be 30. The number of pairwise constraints $N$ for DML is set as 16160 (Actually, we obtain the best result with this value in our experiments. ). The experimental results are plotted in Fig. 1.

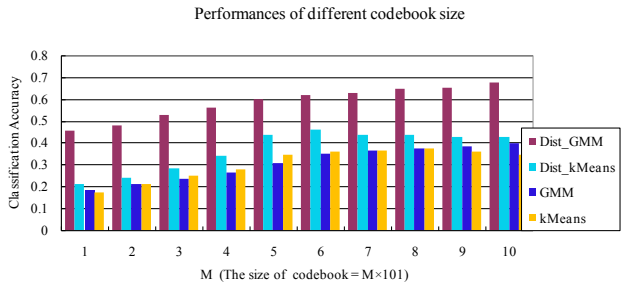


Fig. 1. The effect of different codebook size on classification performance

In Fig. 1, we test four methods based on codebooks: GMM with distance metric, k-means with distance metric, traditional GMM and traditional k-means, all of which use our proposed LM for classification. Since $M$ denotes the number of components for each GMM, the size of codebook is $M$×101, and $M$ takes values from 1 to 10 for testing, meaning that the codebook size takes values from 101 to 1010.

As shown in Fig. 1, after introducing the distance metric, the classification performance is greatly improved, and meanwhile, the result of GMM is better than that of k-means. We obtain the best result using the proposed method. We also find that the performance improves with increasing the codebook size for our method, but for k-means, the performance drops when the codebook size exceeds 600. Finally, we can see that without the distance metric, the classification performance of LM using k-means or GMM based on codebooks is not very well, but with the distance metric for GMM, we get the best performance, which confirms that the combination of DML-based GMM and language model is effective.

### 4.2. Comparison with other classifiers

We compare our method with baselines [25-27], and SVM and Naive Bayes classifiers using BoVW generated by our GMM model. Experimental results under all different numbers of training samples per class are shown in Fig. 2.

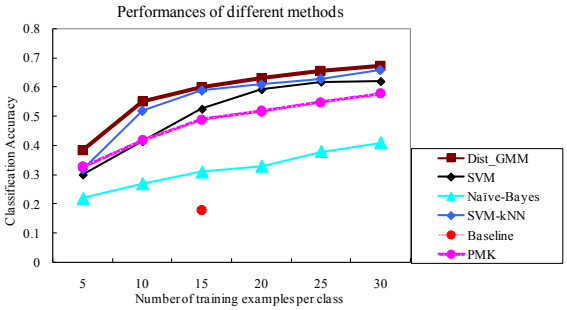


Fig. 2. Performance Comparison with different classifiers

We first use the DML-based GMM to construct histogram features based on BoVW, and then test the SVM and Naive-Bayes classifiers using these features. As shown in Fig. 2, our method is much better than SVM and Naive-Bayes models, which further testifies our idea, namely that the codebook based on GMM trained by DML is very suitable for LM for image classification. Furthermore, our model outperforms several baselines [25-27], because our model achieves the best classification accuracy of 0.67 under 30 training images per class among all these methods.

We also compare our model with state of the art methods. Table 1 shows some state of the art results on Caltech101 dataset.

Table 1. Image classification results with state of the art methods

| Training images | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|
| Dist_GMM | 0.39 | 0.55 | 0.60 | 0.63 | 0.66 | **0.67** |
| Wang et al. [3] | 0.51 | 0.59 | 0.65 | 0.67 | 0.70 | 0.73 |
| NBNN [28] | - | - | 0.65 | - | - | 0.70 |
| Jia et al. [23] | - | - | - | - | - | **0.75** |
| Dist_GMM_SC | | | | | | **0.79** |

The fifth row in Table 1 shows the excellent results [23] in recent research. Our approach does not yet reach approaches to their performance, but when combined with sparse coding (SC) strategy, our method can achieve the best result (As shown by Dist_GMM_SC in Table 1). Furthermore, in Fig. 1, we can see that the trend of the performance of our method increases gradually with the rise of the codebook size.

Finally, we also experiment and learn that the training data of DML have influence on the performance. So, we conclude that there is still much room for our method to be improved.

## 5. CONCLUSION

This paper presents a novel image classification method. We trained GMM with a semantic distance metric for the visual codebook generation, and then construct visual documents using this codebook for a language model classifier. Our method overcomes the drawbacks of the conventional BoVW model which suffers from semantic and spatial information loss. Furthermore, we also learned the advantage of the language model combined with GMM based visual words. The experiments on Caltech101 dataset confirmed the effectiveness of our method. The proposed method achieved the best performance compared with conventional methods. And compared with the state of the art excellent results, our method is also competitive and satisfactory.

In the future, we will explore the more efficient and automatic selecting method of semantic pairwise constraints for DML, and meanwhile we will find more effective visual words sequence for the language model. We believe these two aspects can further improve the image classification performance.

## REFERENCES

[1] Huang Y, Wu Z, Wang L, et al. "Feature coding in image classification: A comprehensive study," *Pattern Analysis and Machine Intelligence*, IEEE Transactions, 36(3):pp. 493-506, 2014.
[2] J. Yang, K. Yu, Y. Gong, T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 1794–1801, 2009.
[3] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, "Locality-constrained linear coding for image classification," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 3360–3367, 2010.
[4] X. Zhou, K. Yu, T. Zhang, T. S. Huang, "Image classification using supervector coding of local image descriptors," *Proc of Computer Vision–ECCV*, Springer, pp. 141–154, 2010.
[5] Zhou N, Fan J. "Jointly learning visually correlated dictionaries for large-scale visual recognition applications," *Pattern Analysis and Machine Intelligence*, IEEE Transactions, 36(4):pp. 715-730, 2014.
[6] F. Jurie, B. Triggs, "Creating efficient codebooks for visual recognition", *Proc of Computer Vision (ICCV)*, IEEE, pp. 604–610, 2005.
[7] J. Wu, J. M. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," *Proc of Computer Vision (ICCV)*, IEEE, pp. 630–637, 2009.
[8] J. C. van Gemert, C. J. Veenman, A. W. Smeulders, J.-M. Geusebroek, "Visual word ambiguity," *Pattern Analysis and Machine Intelligence*, IEEE Transactions, 32(7):pp. 1271–1283, 2010.
[9] Z. Jiang, Z. Lin, L. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *Pattern Analysis and Machine Intelligence*, IEEE Transactions, 35(11):pp. 2651–2664, 2013.
[10] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *Pattern Analysis and Machine Intelligence*, IEEE Transactions, 30 (7):pp. 1243–1256, 2008.
[11] X. Zhou, X. Zhuang, H. Tang, M. Hasegawa-Johnson, T. S. Huang, "Novel gaussianized vector representation for improved natural scene categorization," *Pattern Recognition Letters*, 31(8):pp. 702–708, 2010.
[12] H. Nakayama, T. Harada, Y. Kuniyoshi, "Global gaussian approach for scene categorization using information geometry," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 2336–2343, 2010.
[13] S. N. Parizi, J. G. Oberlin, P. F. Felzenszwalb, "Reconfigurable models for scene recognition," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 2775–2782, 2012.
[14] L. Wu, S. C. Hoi, N. Yu, "Semantics-preserving bag-of-words models and applications," *Image Processing*, IEEE Transactions, 19(7):pp. 1908–1920, 2010.
[15] P. Jain, B. Kulis, K. Grauman, "Fast image search for learned metrics," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 1–8, 2008.
[16] F. Wang, S. Jiang, L. Herranz, Q. Huang, "Improving image distance metric learning by embedding semantic relations," *Advances in Multimedia Information Processing–PCM*, Springer, pp. 424–434, 2012.
[17] Z. Wang, Y. Hu, L.-T. Chia, "Image-to-class distance metric learning for image classification," *Proc of Computer Vision–ECCV*, Springer, pp. 706–719, 2010.
[18] S. Wang, S. Jiang, Q. Huang, Q. Tian, "Multi-feature metric learning with knowledge transfer among semantics and social tagging," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 2240–2247, 2012.
[19] K. Grauman, F. Sha, S. J. Hwang, "Learning a tree of metrics with disjoint visual features," *Advances in Neural Information Processing Systems*, pp. 621–629, 2011.
[20] L. Wu, M. Li, Z. Li, W.-Y. Ma, N. Yu, "Visual language modeling for image classification," *Proc of the international workshop on Workshop on multimedia information retrieval*, ACM, pp. 115–124, 2007.
[21] Penatti O A B, Silva F B, Valle E, et al. "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognition*, 47(2):pp. 705-720, 2014.
[22] L. Fei-Fei, R. Fergus, P. Perona, "One-shot learning of object categories," *Pattern Analysis and Machine Intelligence*, IEEE Transactions, 28(4):pp. 594–611, 2006.
[23] Y. Jia, C. Huang, T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 3370–3377, 2012.
[24] C. D. Manning, P. Raghavan, H. Schu¨tze, *Introduction to information retrieval*, Cambridge University Press Cambridge, 2008.
[25] L. Fei-Fei, R. Fergus, P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, 106 (1):pp. 59–70, 2007.
[26] H. Zhang, A. C. Berg, M. Maire, J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 2126–2136, 2006.
[27] K. Grauman, T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," *Proc of Computer Vision (ICCV)*, IEEE, pp. 1458–1465, 2005.
[28] O. Boiman, E. Shechtman, M. Irani, "In defense of nearest-neighbor based image classification," *Proc of Computer Vision and Pattern Recognition*, IEEE, pp. 1–8, 2008.