Undo The Codebook Bias by Linear Transformation for Visual Applications

Chunjie Zhang[1], Yifan Zhang[2], Shuhui Wang[3], Junbiao Pang[4], Chao Liang[5],
Qingming Huang[1,3], Qi Tian[6]

[1]School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049, Beijing, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Key Lab of Intell. Info. Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[4]College of Computer Science and Technology, Beijing University of Technology, 100124, China
[5]National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, 430072, Wuhan, China
[6]Department of Computer Sciences, University of Texas at San Antonio, TX, 78249, U.S.A

{cjzhang, shwang, qmhuang}@jdl.ac.cn, yfzhang@nlpr.ia.ac.cn,
junbiao_pang@bjut.edu.cn, cliang@whu.edu.cn, qitian@cs.utsa.edu

ABSTRACT

The bag of visual words model (BoW) and its variants have demonstrate their effectiveness for visual applications and have been widely used by researchers. The BoW model first extracts local features and generates the corresponding codebook, the elements of a codebook are viewed as visual words. The local features within each image are then encoded to get the final histogram representation. However, the codebook is dataset dependent and has to be generated for each image dataset. This costs a lot of computational time and weakens the generalization power of the BoW model. To solve these problems, in this paper, we propose to undo the dataset bias by codebook linear transformation. To represent every points within the local feature space using Euclidean distance, the number of bases should be no less than the space dimensions. Hence, each codebook can be viewed as a linear transformation of these bases. In this way, we can transform the pre-learned codebooks for a new dataset. However, not all of the visual words are equally important for the new dataset, it would be more effective if we can make some selection using sparsity constraints and choose the most discriminative visual words for transformation. We propose an alternative optimization algorithm to jointly search for the optimal linear transformation matrixes and the encoding parameters. Image classification experimental results on several image datasets show the effectiveness of the proposed method.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*representation, data structures, and transforms*

Keywords

codebook bias, linear transformation, sparsity

1. INTRODUCTION

The bag of visual words model (BoW)[1] plays a very important role for visual applications (*e.g.* image classification, retrieval and segmentation). Basically, the BoW model can be divided into four components: local feature extraction, codebook generation, local feature encoding and histogram based image representation. It has been widely used on various datasets [2-8] with encouraging visual application results. To consider the spatial information, spatial pyramid matching (SPM)[4] and its variants are also widely used by researchers.

Although proven effective, there is one problem with the BoW model and its variants. The codebook has to be learned for each image dataset separately and the performances of directly using codebooks generated by other datasets are less competitive than using the codebook generated with the corresponding dataset. This is because the state-of-the-art image datasets are collected for particular purposes[9]. To overcome this problem, a lot of works [10-23] have been done. Khosla *et al.* [10] tried to undo the dataset bias by jointly learning the bias vectors and visual words' weights in a discriminative manner. An online domain adaption of cascade classifiers is proposed by Jain and Miller [11]. Kulis *et al.* [12] proposed an asymmetric kernel transformation based object categorization method. The dataset shift problem is systematically analyzed by Candela *et al.* [13]. Zhang *et al.* [14] proposed a non-negative sparse coding based image classification approach. Gopalan *et al.* [15] took an unsupervised approach to adapt the object categorization problem while Zhang *et al.* [16] used a bilinear model for recognition. Saenko *et al.* [17] tried to adapt object model of a particular visual domain to new domain by minimizing the effect of

MM '13, October 21 - 25 2013, Barcelona, Spain
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502141.

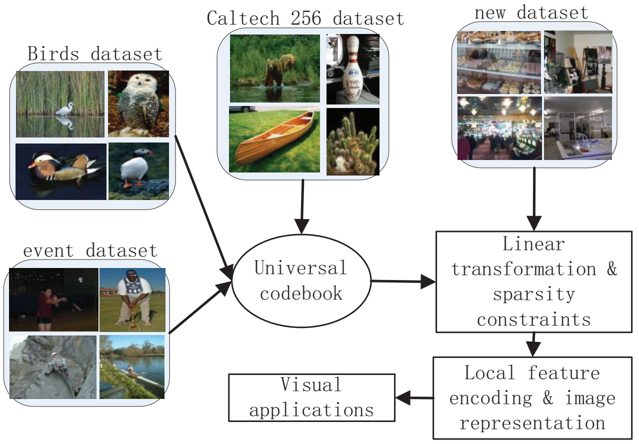

Figure 1: Flowchart of the proposed linear codebook transformation for visual application method.

feature distribution discrepancy While Zhang *et al.* [18] adopt the weak semantic representation. A heterogeneous transfer learning algorithm is proposed by Zhu *et al.* [19] for image classification with good performance. Zhang *et al.* [20] used harr-like transformation of local features. To cope with the lack of training images, Wang *et al.* [21] proposed a dyadic knowledge transfer approach for cross-domain classifications.

All of these methods demonstrate the usefulness of considering the dataset bias for visual applications. However, most of these methods ignore the codebook bias problem with different datasets and only try to adapt the pre-learned classifiers instead. In fact, if we take a close look at the four components of the BoW model, we can find that the codebook is the only component which varies from datasets. The other three components of the BoW model have not such dataset dependence. For example, dense SIFT feature is used for local feature extraction, sparse coding or nearest neighbor assignment is used for local feature encoding and images are represented by visual word histogram. Hence, if we can cope with the codebook bias problem, we will be able to make the BoW model less dataset dependent and improve the performance of visual applications using the BoW representation.

To solve the codebook bias problem, in this paper, we propose a novel linear transformation based codebook adaption method. For the local feature space, the number of bases should be no less than the space dimensions. Hence, we try to view each codebook as a linear transformation of these bases. In this way, we can linearly transform the pre-learned codebooks for a new dataset. However, not all of the visual words are equally important, it is more effective if we can choose the most discriminative visual words for transformation. We use the popular sparsity constraints in this paper for visual word selection. Besides, we also propose an alternative optimization algorithm to jointly search for the optimal linear transformation matrixes and the encoding parameters. To test the effectiveness of the proposed method, we conduct image classification experiments on several image datasets. The results show the effectiveness of codebook transformation for undoing the dataset bias. Figure 1 shows the flowchart of the proposed method.

The rest of this paper is organized as follows. In Section 2 we give the details of the proposed linear codebook transform method for undoing the codebook bias. The experimental results are given in Section 3. Finally, we conclude in Section 4.

2. LINEAR CODEBOOK TRANSFORMATION FOR VISUAL APPLICATIONS

In this section, we will give the details of the proposed linear codebook transform method to undo the codebook bias and apply it to image classification problems.

2.1 Undo the dataset bias by linear codebook transformation

For the local feature space, the number of bases should be no less than the space dimensions. Suppose we have a set of bases $B = [b_1, b_2, ..., b_Q] \in \mathbb{R}^{P\times Q}$ which can completely represent each local feature in this space. P is the dimension of local feature space and Q is the number of bases with $Q > P$. Let $D_1 = [d_1^1, d_1^2, ..., d_1^M] \in \mathbb{R}^{\mathbb{P}\times\mathbb{M}}$ be a codebook generated using a particular dataset where M is the number of visual words. Since each visual word in codebook D_1 can be viewed as a point in the local feature space, each element of D_1 can be linearly represented by B as:

$$d_1^i = a_1^1 b_1 + a_1^2 b_2 + ... + a_1^Q b_Q, \forall i = 1, ..., Q \quad (1)$$

This can be rewritten in a matrix form as:

$$D_1 = A_1 B \quad (2)$$

with A_1 is the corresponding linear transformation matrix. In this way, we can generate a codebook D_1 by linearly combine the bases of local feature space. This can also be written as:

$$B = A_1^+ D_1 \quad (3)$$

Where A_1^+ is the psedoinverse of matrix A_1. Similarly, we can generate a codebook D_2 as:

$$D_2 = A_2 B = A_2 A_1^+ D_1 \quad (4)$$

Let $A = A_2 A_1^+$, Eq.4 can be rewritten as:

$$D_2 = A D_1 \quad (5)$$

Suppose we have learnt the codebook D_1 for dataset 1, to generate the codebook D_2 for dataset 2, all we need to do is to find the corresponding transformation matrix A. If the transformation matrix A has been learnt, we can use the corresponding codebook D_2 for local feature encoding. We use the sparse coding technique [24] in this paper as it has been shown very effective for encoding local features. Let $x \in \mathbb{R}^{P\times 1}$ be the local feature to be encoded, α is the corresponding sparse coding parameter with λ is the parameter which controls the sparsity of α as:

$$min_{\alpha,D_2} \| x - \alpha^T D_2 \|^2 + \lambda \| \alpha \|_1 \quad (6)$$

This can be optimized over α and A as:

$$min_{\alpha,A} \| x - \alpha^T A D_1 \|^2 + \lambda \| \alpha \|_1 \quad (7)$$

This problem can be solved efficiently by alternatively optimizing over α/A while keeping the other fixed. When α is fixed, Eq.7 equals to solving the following optimization problem as:

$$min_A \| x - \alpha^T A D_1 \|^2 \quad (8)$$

When A is fixed, Eq.7 equals to solving the following optimization problem as:

$$min_{\alpha} \parallel x - \alpha^T A D_1 \parallel^2 + \lambda \parallel \alpha \parallel_1 \tag{9}$$

Since D_1 is pre-learned and fixed, let $D = AD_1$, Eq.9 can be rewritten as:

$$min_{\alpha} \parallel x - \alpha^T D \parallel^2 + \lambda \parallel \alpha \parallel_1 \tag{10}$$

Eq.8 and Eq.10 can be effectively optimized by the feature-sign search algorithm and the Lagrange dual algorithm proposed in [24]. In this way, we can transform the codebook D_1 generated using dataset 1 to dataset 2 accordingly. However, the transformation of only one codebook is often too weak, especially when the two image datasets are quite different. It would be more effective if we can transform a number of codebooks for an unseen dataset.

Formally, suppose we have M pre-learned codebooks generated using the corresponding image datasets. To encode local feature x, the optimization problem can be written similarly as:

$$min_{\alpha_i, A_i, i=1,2,...,M} \parallel x - \sum_{i=1}^{M} \alpha_i^T A_i D_i \parallel^2 + \lambda_i \sum_{i=1}^{M} \parallel \alpha_i \parallel_1 \tag{11}$$

Where λ_i is the sparsity constraint parameter for the $i-th$ dataset, α_i is the corresponding encoding parameter. Let $\beta = [\alpha_1; \alpha_2; ...; \alpha_M]$, $E = [D_1; D_2; ...; D_M]$ and $C = diag A_1, A_2, ..., A_M$, Eq.11 can be rewritten as:

$$min_{\beta, C} \parallel x - \beta^T C E \parallel^2 + \lambda \parallel \beta \parallel_1 \tag{12}$$

This problem can be solved similarly as Eq.7 by alternatively optimizing over β and C.

2.2 Max pooling based image representation for visual applications

After learning the corresponding linear transformation matrix C, we can use it to encode local features by fixing C. We encode each local feature individually. We follow the popular max pooling scheme [14, 16, 18, 20, 22, 23, 25] to extract information from local features for image representation. The max pooling is proven effective when combined with sparse coding for image representation. Besides, to combine the spatial information of local features, we adopt the spatial pyramid matching (SPM) technique [4]. We use the first three pyramids as $2^L \times 2^L, L = 0, 1, 2$.

To test the effectiveness of the proposed linear codebook transformation method for undoing the dataset bias, we conduct image classification performances. This is achieved by training a set of classifiers. We use the one-vs-all linear SVM classifier.

3. EXPERIMENTS

To evaluate the effectiveness of the proposed linear codebook transformation method, we conduct image classification performance on several public image datasets: the Bird dataset[2], the Butterfly dataset[3], the Scene-15 dataset[4], the Event dataset[5], the Indoor dataset[6], the Corel-5K dataset[7] and the Caltech-256 dataset[8]. We densely extract SIFT features of size 16×16 pixels with an overlap of 6 pixels. For the seven datasets, we randomly choose 50, 16, 100, 70, 80, 50 and 30 images per class for the corresponding image dataset. This process is repeated for five times to get reliable results. The codebook size for each dataset is set to 1,024. We use the classification rate as the performance measurement method.

We give the performance comparison of the proposed linear codebook transfer for undoing the dataset bias algorithm on the seven image datasets in Table 1. The horizontal row indicates the dataset that the codebook is generated while the vertical column indicates the dataset that the classification is performed. We also give the performance of the proposed linear codebook transfer algorithm on the corresponding vertical column by transfering the codebook generated by the other six datasets.

We can see from Table 1 that the codebook generated by one particular image dataset achieves the best classification performance on the corresponding dataset. This is because of the manual selection of images for particular purposes [9]. However, we can see from Table 1 that we can achieve better results by transfering the codebooks instead of directly using the codebooks generated by other datasets. In fact, the codebook generated by the corresponding image dataset is the upper performance bound of the proposed codebook transfer algorithm.

Besides, the relative improvement of the proposed codebook transfer algorithm varies over image datasets. For example, the proposed method achieves equal performance on the Butterfly and Corel-5K datasets while performs 3/2 percent less on the Caltech-256/Indoor dataset compared with the codebook generated by the corresponding datasets. We believe this is because the difficulties of these datasets are different. The Caltech-256 dataset and the Indoor dataset are more difficult to classify that the Corel-5K dataset and the Butterfly dataset. This is not only because of the increased number of image classes but also because of the large intra and inter class variations.

On analyzing the details of the classification performance, we can have two conclusions. First, Compared with the codebook generated by the corresponding dataset, the use of other datasets generated codebooks perform better on similar image classes than on dissimilar image classes. For example, the Scene-15 dataset can be roughly divided into the indoor class and the outdoor class. When using the Indoor dataset generated codebook for classification, the performances are comparable or a little less that the Scene-15 generated codebook on the indoor class (*e.g.* kitchen, livingroom, store). However, for the outdoor class (e.g. highway/mountain decreases by 4/3 percent respectively). Fortunately, the proposed codebook transfer method can alleviate this problem by transfer the elements of datasets with similar image classes for better image representation. We can achieve comparable classification rates by codebook transformation (*e.g.* the Scene-15 dataset and the Corel-5K dataset). These results prove the effectiveness of transfering codebook for undoing the dataset bias and improve the classification performance.

4. CONCLUSION

In this paper, we proposed a novel linear codebook transformation method to undo the codebook bias. This is achieved by linearly transform the pre-learned codebooks for new visual applications. We also impose sparisity constraints for discriminative visual words transformation. An alternative optimization algorithm is proposed to jointly learn the optimal transformation matrix and encoding parameters. Experimental results on seven public datasets prove the effectiveness of the proposed method.

5. ACKNOWLEDGEMENT

This work is supported in part by National Basic Research Program of China (973 Program): 2012CB316400; the President Fund of UCAS; the Open Project Program of the National Laboratory of Pattern Recognition (NLPR); China Postdoctoral Science Founda-

Table 1: Mean classification rates on the seven image datasets: the Bird dataset, the Butterfly dataset, the Scene-15 dataset, the Event dataset, the Indoor dataset, the Corel-5K dataset and the Caltech-256 dataset. The horizontal row indicates the dataset that the codebook is generated while the vertical column indicates the dataset that the classification is performed. We also give the performance of the proposed linear codebook transfer algorithm by transfering the codebook generated by the other six datasets on the corresponding vertical column.

datasets	Bird	Butterfly	Scene-15	Event	Indoor	Corel-5K	Caltech-256
Bird	**0.83± 0.07**	0.72 ± 0.09	0.78 ± 0.06	0.79 ± 0.07	0.39 ± 0.08	0.61 ± 0.04	0.29 ± 0.06
Butterfly	0.75 ± 0.08	**0.72± 0.08**	0.77 ± 0.06	0.78 ± 0.07	0.38 ± 0.06	0.60 ± 0.04	0.29 ± 0.05
Scene-15	0.72 ± 0.06	0.69 ± 0.08	**0.79± 0.05**	0.78 ± 0.08	0.40 ± 0.07	0.62 ± 0.05	0.28 ± 0.06
Event	0.73 ± 0.06	0.73 ± 0.07	0.74 ± 0.07	**0.81± 0.07**	0.41 ± 0.07	0.61 ± 0.03	0.31 ± 0.06
Indoor	0.70 ± 0.08	0.72 ± 1.01	0.77 ± 0.05	0.79 ± 0.08	**0.43± 0.06**	0.62 ± 0.04	0.32 ± 0.07
Corel-5K	0.72 ± 0.09	0.70 ± 0.09	0.76 ± 0.06	0.78 ± 0.08	0.39 ± 0.08	**0.67± 0.05**	0.31 ± 0.05
Caltech-256	0.71 ± 0.08	0.71 ± 0.08	0.75 ± 0.05	0.79 ± 0.09	0.40 ± 0.05	0.64 ± 0.04	**0.38± 0.06**
codebook transfer	**0.81± 0.07**	**0.72± 0.08**	**0.78± 0.05**	**0.80± 0.08**	**0.41± 0.07**	**0.67± 0.04**	**0.35± 0.06**

tion: 2012M520434, 2013T60156, 2013M530350, 2013M530739; National Natural Science Foundation of China: 61025011, 61202325, 61202234; Beijing Municipal Natural Science Foundation of China No.4132010. This work is supported in part to Dr. Qi Tian by ARO grant W911BF-12-1-0057, NSF IIS 1052851, Faculty Research Awards by Google, FXPAL, and NEC Laboratories of America and 2012 UTSA START-R Research Award respectively. This work is supported in part by NSFC 61128007.

6. REFERENCES

[1] J. Sivic and A. Zisserman, Video google: A text retrieval approach to object matching in videos, *ICCV*, pp.1470-1477, UK, 2003.

[2] S. Lazebnik, C. Schmid, and J. Ponce, A maximum entropy framework for part-based texture and object recognition, *ICCV*, pp.832-838, China, 2005.

[3] S. Lazebnik, C. Schmid, and J. Ponce, Semi-local affine parts for object recognition, *BMVC*, pp. 959-968, 2004.

[4] S. Lazebnik, C. Schmid, and J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *CVPR*, pp.2169-2178, USA, 2006.

[5] L. Li, and Li Fei-Fei, What, where and who? Classifying event by scene and object recognition, *ICCV*, pp.1-8, 2007.

[6] A. Quattoni and A. Torralba, Recognizing indoor scenes, *CVPR*, pp.413-420, USA, 2009.

[7] G. Liu, Z. Li, L. Zhang, and Y. Xu, Image retrieval based on micro-structure descriptor, *Pattern Recognition*, 44(9):2123-2133, 2011.

[8] G. Griffin, A. Holub, and P. Perona, Caltech-256 object category dataset, Technical Report, CalTech, 2007.

[9] A. Torralba and A. Efros, Unbiased look at dataset bias, *CVPR*, pp.1521-1528, USA, 2011.

[10] A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, and A. Torralba, Undoing the damage of dataset bias, *ECCV*, pp.158-171, 2012.

[11] V. Jain and E. Miller, Online domain adaption of a pre-trained cascade of classifiers, *CVPR*, pp.577-584, USA, 2011.

[12] B. Kulis, K. Saenko, and T. Darrell, What you saw is not what you get: Domain adaptation using asymmetric kernel transforms, *CVPR*, pp.1785-1792, USA, 2011.

[13] J. Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, *Dataset shift in machine learning*, MIT Press, 2009.

[14] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu and S. Ma, Image classification by non-negative sparse coding, low-rank and sparse decomposition, *CVPR*, 2011, pages 1673-1680.

[15] R. Gopalan, R. Li, and R. Chellappa, Domain adaptation for object recognition: An unsupervised approach, *ICCV*, pp.999-1006, 2011.

[16] C. Zhang, J. Liu, Q. Tian, Y. Han, H. Lu, S. Ma, A Boosting Sparsity-Constrained Bilinear Model for Object Recognition, *IEEE Multimedia*, 2012, 19(2):58-68.

[17] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, Adapting visual category models to new domains, *ECCV*, pp.213-226, 2010.

[18] C. Zhang, J. Liu, Q. Tian, C. Liang, Q. Huang, Beyond Visual Features: A Weak Semantic Image Representation Using Exemplar Classifiers for Classification, *Neurocomputing*, DOI:10.1016/j.neucom.2012.07.056.

[19] Y. Zhu, Y. Chen, Z. Lu, S. Pan, G. Xue, Y. Yu and Q. Yang, Heterogeneous transfer learning for image classification, *AAAI*, USA, 2011.

[20] C. Zhang, J. Liu, Q. Tian, C. Liang, Q. Huang, Image Classification Using Harr-like Transformation of Local Features with Coding Residuals, *Signal Processing*, 2012, 93(8):2111-2118.

[21] H. Wang, F. Nie, H. Huang, and C. Ding, Dyadic transfer learning for cross-domain image classification, *ICCV*, pp.551-556, 2011.

[22] C. Zhang, S. Wang, Q. Huang, J. Liu, C. Liang, Q. Tian, Image Classification Using Spatial Pyramid Robust Sparse Coding, *Pattern Recognition letters*, 34(9):1046-1052.

[23] C. Zhang, S. Wang, Q. Huang, C. Liang, J. Liu, Q. Tian, Laplacian affine sparse coding with tilt and orientation consistency for image classification, *Journal of Visual Communication and Image Representation*, 24(7), pp.786-793, 2013.

[24] H. Lee, A. Battle, R. Raina, and A. Ng, Efficient sparse coding algorithms, *NIPS*, 2007.

[25] C. Zhang, J. Liu, J. Wang, Q. Tian, C. Xu, H. Lu, S. Ma, Image Classification Using Spatial Pyramid Coding and Visual Word Reweighting, *ACCV* 2010, pp. 239-249.