Neurocomputing 77 (2012) 281–288

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Letters

# A unified supervised codebook learning framework for classification

Congyan Lang [a,*], Songhe Feng [a,c], Bing Cheng [b], Bingbing Ni [b], Shuicheng Yan [b]

[a] Department of Computer Science and Engineering, Beijing Jiaotong University, Beijing 100044, China
[b] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[c] Beijing Key Lab of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China

ARTICLE INFO

*Article history:*
Received 29 May 2011
Received in revised form
4 September 2011
Accepted 4 September 2011
Communicated by Tao Mei
Available online 24 September 2011

*Keywords:*
Visual dictionary
Bag-of-words
K-means
Supervised learning

ABSTRACT

In this paper, we investigate a *discriminative* visual dictionary learning method for boosting the classification performance. Tied to the K-means clustering philosophy, those popular algorithms for visual dictionary learning cannot guarantee the well-separation of the normalized visual word frequency vectors from distinctive classes or large label distances. The rationale of this work is to harness sample label information for learning visual dictionary in a supervised manner, and this target is then formulated as an objective function, where each sample element, *e.g.*, SIFT descriptor, is expected to be close to its assigned visual word, and at the same time the normalized aggregative visual word frequency vectors are expected to possess the property that kindred samples shall be close to each other while inhomogeneous samples shall be far away. By relaxing the hard binary constraints to soft nonnegative ones, a multiplicative nonnegative update procedure is proposed to optimize the objective function along with theoretic convergence proof. Extensive experiments on classification tasks (*i.e.*, natural scene and sports event classifications) all demonstrate the superiority of this proposed framework over conventional clustering based visual dictionary learning.



## 1. Introduction

Bag-of-words (BoW) is a state-of-the-art approach for modeling the global statistics of the local visual features for image [3] or video [13] representation. Typically, a set of visual words, namely the so-called visual dictionary, are learned from the training samples, and then each image or video is expressed as a bag of these words, finally the normalized occurrence frequencies of these visual words are used for data representation. As a rule of thumb, these visual words were learned by unsupervised clustering approaches, *e.g.*, K-means [5], driven by the philosophy that the visual words should be the centers of data clusters. The resulting histogram representations are extensively used in visual classification tasks [3,4,15–17] as well as visual regression tasks [19].

However, when given label information of the training samples, the above philosophy becomes not appealing due to its incapability in guaranteeing discriminating power for the normalized visual word frequency vectors. This behavior results in a non-optimal solution for the codebook, since in classification tasks, the histogram representation should preserve the property that: samples from different classes should lie far away in the feature space while samples from the same class should lie close to each other. A natural problem to study is how to harness the label information for boosting the visual classification capability of the final normalized aggregative visual word frequency representation. Fig. 1 shows a toy problem, where supervised visual dictionary boosts up the differentiability of the samples from different classes compared to the results from K-means based visual dictionary.

Several approaches have been proposed to address/partially address the above problem. Winn et al. [17] propose a method based on pair-wise word merging to reduce the large vocabulary structure. Moosmann et al. [12] propose to use random forests to construct the visual word vocabulary in a supervised manner. Ning et al. [14] develop a supervised visual dictionary learning method, where a postprocessing step is developed to improve the discriminative power of the derived visual dictionary from K-means approach. These methods typically separate the process of visual codebook generation from the process of classifier training. However, the capability of this postprocessing is limited due to the fact that some useful information for classification may have been lost in calculating the visual word frequencies and is not restorable. Yang et al. [20] propose a discriminative visual codebook generation method based on a two phase training method. By introducing a set of additional representation information *i.e.*, *visual bits* as well as a classifier for each category, their learning algorithm improves the discriminative performance by iterating the optimization in terms of both above aspects. More recently, Lazebnik and Raginsky [7] present a method for supervised learning of quantizer codebook based on information loss minimization. They develop a alternating minimization procedure

* Corresponding author.
*E-mail addresses:* cylang@bjtu.edu.cn (C. Lang), elev44@nus.edu.cn (S. Feng), g0800415@nus.edu.sg (B. Cheng), g0501096@nus.edu.sg (B. Ni), eleyans@nus.edu.sg (S. Yan).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.09.010

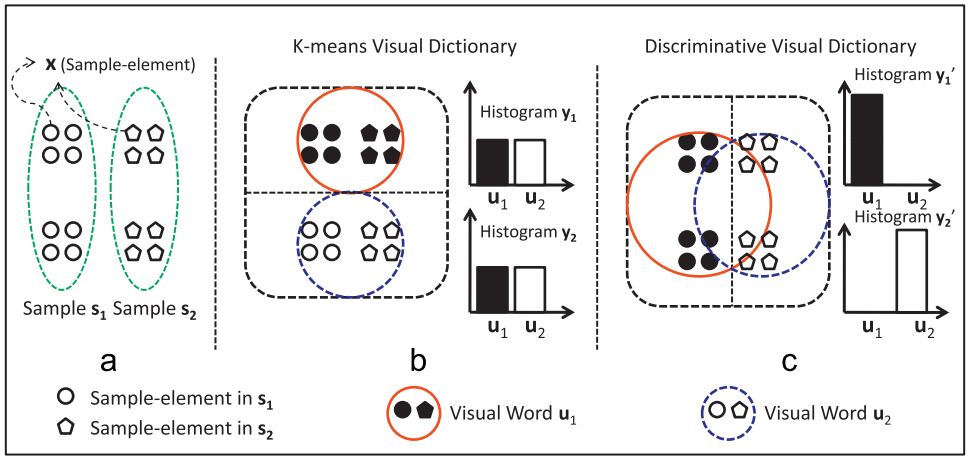


**Fig. 1.** A toy problem with two samples and two visual words. (a) Two samples $\mathbf{s}_1$ and $\mathbf{s}_2$ are from two different classes, and each is expressed with a bag of points. (b) The visual word frequency histograms $\mathbf{y}_1$ and $\mathbf{y}_2$ based on visual dictionary learnt from K-means for data $\mathbf{s}_1$ and $\mathbf{s}_2$, where $\mathbf{y}_1$ and $\mathbf{y}_2$ are not differentiable. (c) The visual word frequency histograms $\mathbf{y}_1'$ and $\mathbf{y}_2'$ based on the discriminative visual dictionary for data $\mathbf{s}_1$ and $\mathbf{s}_2$, where $\mathbf{y}_1'$ and $\mathbf{y}_2'$ are differentiable.

for simultaneously quantizing the continuous input feature vector and approximating the quantizer index of the training samples according to the posterior class label distributions. The limitation of the above methods is that they could only deal with the classification problem and could not be extended to cope with the regression problem based on histogram representations.

To alleviate this problem, in this paper, we present a more general discriminative visual dictionary (DVD) learning framework to strengthen visual dictionary by harnessing the label information of the training samples, which could be applied on classification problems. A novel objective function is proposed by unifying the dual targets, namely, the sample element, *e.g.*, SIFT descriptor, is expected to be close to its assigned visual word, and the final normalized visual word frequency vectors are expected to possess the property that kindred samples shall be close to each other while inhomogeneous samples shall be faraway. By relaxing the hard binary constraints to soft nonnegative ones, the proposed optimization problem can be effectively solved by nonnegative multiplicative update rules, with theoretically provable convergence. Finally the normalized visual word frequency vector for new sample is derived with kernel regression approach. Extensive experiments on natural scene and sports event classifications all demonstrate the encouraging improvements in visual classification performance gained from our proposed framework for learning discriminative visual dictionary.

This paper is organized as follows. In Sections 2 and 3, we give the detailed description of our unified discriminative visual dictionary learning framework. The experimental results of the classification tasks are presented in Section 4, respectively. Section 5 concludes the paper.

## 2. Problem formulation

Before formally introducing the math formulation for the unified discriminative visual dictionary learning framework, we give the terminologies used afterwards. Let $X=[\mathbf{x}_1,\ldots,\mathbf{x}_N]\in\mathbb{R}^{D\times N}$ denote the extracted sample elements, *e.g.*, SIFT descriptors [10], from all training samples. Then each sample element $\mathbf{x}_i$ is represented as a $D$-dimensional vector, and $N$ is the total number of extracted sample elements. The set of labeled training samples (*e.g.*, images and videos) are denoted as $\mathcal{S}=\{\mathbf{s}_1,\ldots,\mathbf{s}_{N_s}\}$, where $N_s$ is the number of training samples. The sample elements in $X$ and belonging to the sample $\mathbf{s}_i$ are denoted by $\mathcal{D}_{\mathbf{s}_i}$, where the number of sample elements from $\mathbf{s}_i$ is $N_{s_i}$, *i.e.*, $N_{s_i}=\left|\mathcal{D}_{\mathbf{s}_i}\right|$. For the case of visual classification tasks, each training sample $\mathbf{s}_i$ is labeled as $c_i\in\{1,\ldots,N_c\}$, where $N_c$ is the number of sample classes.

The goal of the discriminative visual dictionary learning framework is two-fold. On one hand, each visual word is expected to characterize the common properties shared by the sample elements belonging to this visual word, and on the other hand, the final normalized aggregative visual word frequency vector for each sample is expected to be good at distinguishing samples of different classes or far away in the label space, and consequently boot the visual classification performance.

### 2.1. Objective for sample-element compactness

Assume a set of $K$ visual words $U=[\mathbf{u}_1,\ldots,\mathbf{u}_K]$ are to be learnt for fixed-length image or video representation. In classical K-means formulation, these visual words correspond to the cluster centers. Mathematically, the visual word assignment information for the sample element $\mathbf{x}_i$ can be represented as a $K$-dimensional binary vector $\mathbf{v}_i$, the nonzero item of $\mathbf{v}_i$ indicates the index of the corresponding visual word. Then the target of K-means approach is to minimize the sum of distances between the data elements and their assigned visual words, namely,

$$\min_{U,V}\{Q_R(U,V)=\|X-UV^T\|_F^2\},\quad \text{s.t. } V_{ij}\in\{0,1\},$$

where $V=[\mathbf{v}_1,\ldots,\mathbf{v}_N]^T$ and $\|\cdot\|_F$ is the Frobenius norm of a matrix. In this work, the binary constraint on $V$ is relaxed, and as $\mathbf{v}_i$ shall constitute a normalized aggregative visual word frequency vector along with all other sample elements within a sample, the following constraint is imposed instead,

$$\sum_{j=1}^{K}V_{ij}=1,\ \forall i,\quad V_{ij}\geq 0,\ \forall i,j. \tag{1}$$

From the probabilistic point of view, the vector $\mathbf{v}_i$ encodes the probabilities of the sample element $\mathbf{x}_i$ belonging to different visual words. At the same time, the objective function could also be explained from the algebraic point of view, and $U\mathbf{v}_i$ could be regarded as the nonnegative reconstruction of the sample element $\mathbf{x}_i$ based on the visual words $U$. Another observation is that the sample elements $X$, obtained from many image/video feature descriptors are often nonnegative. These properties naturally lead this problem to be a special constrained nonnegative matrix