Pattern Recognition 47 (2014) 899–913

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning group-based dictionaries for discriminative image representation

Hao Lei [a], Kuizhi Mei [a,*], Nanning Zheng [a], Peixiang Dong [a], Ning Zhou [c], Jianping Fan [b,c]

[a] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, PR China
[b] School of Information Science and Technology, Northwest University, Xi'an 710069, PR China
[c] Department of Computer Science, University of North Carolina, Charlotte, NC 28223 USA

ARTICLE INFO

Article history:
Received 15 November 2012
Received in revised form
17 July 2013
Accepted 21 July 2013
Available online 7 August 2013

Keywords:
Group-based dictionary learning
Discriminative image representation
Bag-of-visual-words
Structural learning
Image classification

ABSTRACT

Dictionary learning is a critical issue for achieving discriminative image representation in many computer vision tasks such as object detection and image classification. In this paper, a new algorithm is developed for learning discriminative group-based dictionaries, where the inter-concept (category) visual correlations are leveraged to enhance both the reconstruction quality and the discrimination power of the group-based discriminative dictionaries. A visual concept network is first constructed for determining the groups of visually similar object classes and image concepts automatically. For each group of such visually similar object classes and image concepts, a group-based dictionary is learned for achieving discriminative image representation. A structural learning approach is developed to take advantage of our group-based discriminative dictionaries for classifier training and image classification. The effectiveness and the discrimination power of our group-based discriminative dictionaries have been evaluated on multiple popular visual benchmarks.

1. Introduction

Image content representation is a basic but crucial issue for many computer vision tasks. Intuitively, we can directly extract low-level visual features from images for visual content representation, such as color, texture, and shape. Recently, inspired by the success of bag-of-words for document representation in the domain of text information retrieval, Bag-of-visual-words (BoW) has been widely used for image content representation, which plays an important role in image categorization [1,2], object recognition [3,4], and image retrieval [5,6].

For BoW-based image content representation, an image can be represented as a histogram of the frequencies of "visual words" just as the words in a document representation. However, there does not exist a simple way to obtain effective "visual words" for image content representation. A set of "visual words" (i.e., visual dictionary) should be carefully learnt from the raw visual features for large amounts of training images. Many algorithms have been proposed to learn a universal visual dictionary [3,7–11] for all the object classes and image concepts. We refer to such dictionary learning algorithms as universal dictionary learning approach, where the same bases (the same set of visual words in the universal dictionary) are used to obtain the BoW histograms for all the images in the database. It is worth noting that the images from different object classes and image concepts may have diverse visual properties, thus such universal dictionary may not be optimum and cannot ensure good reconstruction quality and high discrimination power for visual recognition.

Many other algorithms have been proposed to learn an individual visual dictionary (class-specific visual dictionary) [12–16] for each object class or image concept. We refer to such dictionary learning algorithms as individual dictionary learning approach, where a test image can be represented as a set of BoW histograms and each of these BoW histograms corresponds to one class-specific visual dictionary. Learning a large number of class-specific dictionaries can achieve better reconstruction quality and higher discrimination power for visual recognition, but the problem of higher computational complexity may seriously limit its scalability.

When a large number of object classes (i.e., image semantics are interpreted by the visual content of object regions) and image concepts (i.e., image semantics are interpreted by the visual content of entire images) come into view, some of them are strongly inter-related (visually similar) because their relevant images may share some similar or even common visual properties (i.e., strong inter-concept visual correlations) [17–19]. For example, the relevant images for the inter-related image concepts, such as "sky", "cloud", and "wave" may share some similar visual properties and look visually similar. Because of huge inter-concept visual correlations, it is not reasonable for the individual dictionary

* Corresponding author. Tel./fax: +86 29 82668672.
E-mail address: meikuizhi@mail.xjtu.edu.cn (K. Mei).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.07.016

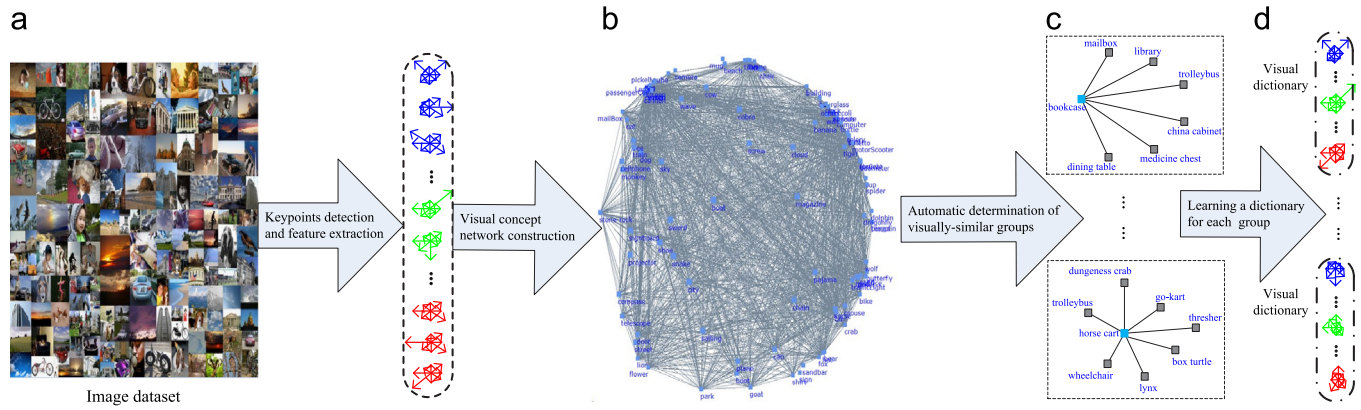

Fig. 1. Flowchart of our group-based algorithm: (a) feature extraction and image representation, (b) visual concept network construction, (c) automatical determination of the groups of visually similar object classes and image concepts, and (d) learning a discriminative dictionary for each group.

learning approach to completely isolate such visually similar object classes and image concepts and learn their inter-related class-specific dictionaries independently [64]. On the other hand, the universal dictionary learning approach cannot learn discriminative dictionaries for distinguishing the visually similar object classes and image concepts effectively.

In this paper, a new algorithm is developed to learn the group-based discriminative dictionaries, where a discriminative group-based dictionary is learned for the visually similar object classes and image concepts in the same group. By leveraging the inter-concept visual correlations for discriminative dictionaries learning, our group-based discriminative dictionaries can achieve better reconstruction quality and higher discrimination power for visual recognition. The brief flowchart of our new algorithm for learning such group-based discriminative dictionaries is illustrated in Fig. 1. In our proposed algorithm, how to characterize the inter-concept visual correlations among the object classes and image concepts is an important issue. Some previous works [20–22] have used the conceptual similarity to characterize the inter-concept correlations, such as WordNet distance [23] and Google distance [24], which characterize the inter-concept relationships by leveraging the knowledge of WordNet ontology [25] and the contextual information on the web pages, respectively. In this paper, a new method is used to characterize the inter-concept correlations by using their visual similarity relationships in the visual feature space. The contributions of this paper can be summarized as follows:

(a) A visual concept network is constructed to characterize the inter-concept visual correlations directly in the visual feature space rather than in the semantic space (i.e., label space), which can be used to determine the groups of visually similar object classes and image concepts automatically.
(b) A new algorithm is developed to leverage the inter-concept visual correlations for learning the group-based discriminative dictionaries with better reconstruction quality and higher discrimination power for visual recognition.
(c) A structural learning approach is developed to take advantage of our group-based discriminative dictionaries for classifier training and image classification.

The rest of this paper is organized as follows. Section 2 reviews the related work briefly. Section 3 presents our work on determining the groups of visually similar object classes and image concepts, where a visual concept network is constructed to characterize the inter-concept visual correlations precisely and explicitly. Section 4 describes our approach for learning the group-based discriminative dictionaries in more details. Section 5 presents a structural learning approach to leverage our group-based discriminative dictionaries for classifier training and image classification. Section 6 presents our experimental results, which are compared with other existing approaches on reconstruction error and discrimination power. We conclude this paper in Section 7.

2. Related work

Image content representation is an attractive topic, and it is related to many research areas (e.g., feature extraction, dimension reduction, and visual signature generation). Comprehensive reviews could be found in the literature [5,6,26,27]. This paper focuses on dictionary learning for supporting BoW-based image content representation, thus we mainly review some related work on BoW-based image content representation and dictionary learning.

Because of its effectiveness and flexibility, the bag-of-visual-words approach has become one of most popular methods for image content representation. Using BoW for image content representation generally consists of the following steps: (a) local image patch extraction; (b) feature detection and patch description; (c) visual dictionary (also called vocabulary) learning; and (d) frequency (histogram) of visual words computation. In the patch extraction step, some salient "interesting" local patches are extracted from the images, and many approaches have been suggested: patch extraction based on segmentation [28], at the key points [1,29], by using regular grid [3,30], or at random [2,31]. There are several descriptors for these local patches such as shape context, complex filters, SIFT (Scale Invariant Feature Transform) [32], PCA-SIFT [33] and so on. During these descriptors, the SIFT-based approach has obtained outstanding performance referring to [26]. Visual dictionary learning is an important step, where the feature vectors are generally clustered by using K-means [1], Gaussian Mixture Models (GMMs) [34,35], mean-shift [31], etc. The centriods of these clusters are treated as the visual words, and the number of clusters determines the size of visual dictionary, which can vary from hundreds to over tens of thousands. Once the visual dictionary is generated, the feature vectors extracted from an image can be represented by their closest visual words in the visual dictionary. This process can be demonstrated as a brief formula: assume D is a dictionary that has m visual words $\{d_1,\ldots,d_m\}$, $X=\{x_1,\ldots,x_n\}$ is a set of feature vectors extracted from an image, the visual content of the image can be represented by the visual words as

$$\mathcal{I}(x_i)=d_j,\ \forall k(k\neq j)\ \mathcal{R}(x_i,d_j)\leq\mathcal{R}(x_i,d_k) \tag{1}$$