Information Fusion 36 (2017) 275–283

Contents lists available at ScienceDirect

Information Fusion

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

Visual feature coding based on heterogeneous structure fusion for image classification

Guangfeng Lin*, Caixia Fan, Hong Zhu, Yalin Miu, Xiaobing Kang

Information Science Department, Xi'an University of Technology, Xi'an, 5 South Jinhua Road, Shaanxi Province 710048, PR China

ARTICLE INFO

Article history:
Received 26 April 2016
Revised 5 December 2016
Accepted 27 December 2016
Available online 28 December 2016

Keywords:
Words structure
Images structure
Heterogeneous structure fusion
Image classification

ABSTRACT

The relationship between visual words and local feature (words structure) or the distribution among images (images structure) is important in feature encoding to approximate the intrinsically discriminative structure of images in the Bag-of-Words (BoW) model. However, in recently most methods, the intrinsic invariance in intra-class images is difficultly captured using words structure or images structure for large variability image classification. To overcome this limitation, we propose a local visual feature coding based on heterogeneous structure fusion (LVFC-HSF) that explores the nonlinear relationship between words structure and images structure in feature space, as follows. First, we utilize high-order topology to describe the dependence of the visual words, and use the distance measurement based on the local feature to represent the distribution of images. Then, we construct the unitedly optimal framework according to the relevance between words structure and images structure to solve the projection matrix of local feature and the weight coefficient, which can exploit the nonlinear relationship of heterogeneous structure to balance their interaction. Finally, we adopt the improving fisher kernel(IFK) to fit the distribution of the projected features for obtaining the image feature. The experimental results on ORL, 15 Scenes, Caltech 101 and Caltech 256 demonstrate that heterogeneous structure fusion significantly enhances the intrinsic structure construction, and consequently improves the classification performance in these data sets.

1. Introduction

Image classification is a fundamental and challenging problem in computer vision. Complex image content (huge amount of images, cluttered background, the differences of image's scale, rotation, transformation and viewpoints) leads to difficultly recognize the correct category of an image. To settle these issues, feature extraction (for example histogram of oriented gradient (HOG) [1] and invariant feature transform (SIFT) [2]), feature encoding (such as Bag-of-Words (BOW) [3,4], sparse coding [5], hierarchical feature coding [6] and deep hierarchical feature coding [7]), feature learning (such as subspace learning [8], metric learning [9] and deep learning [10–12]), and classifier construction (such as support vector machine (SVM) [13] and nearest neighbor (NN) classifier [14]) have been proposed for image classification. In recent years, feature encoding and feature learning have become research focus because of the key role of image representation.

Feature encoding usually is unsupervised method to capture the intrinsic structure for image representation, and feature learning generally involves some supervised methods to use the class label information for image description. The intrinsic structure of images can show the natural characteristic of images, and is important to image classification. In fact, the intrinsic structure of images can be described by heterogeneous structure. In the past ten years, the Bag-of-Words model has been successfully used because of trying to find the intrinsic structure of images components (pixel, feature, image) [4], while the deep learning has obtained a great achievement in large-scale application in terms of the powerful learning ability of the deep networks [10–12]. However, heterogeneous structure fusion mechanism is neglected and is not clear for feature encoding. Therefore, this paper mainly focuses on how to consider and approach the intrinsic structure by heterogeneous structure fusion for image classification in the BoW model. The further understanding of the intrinsic structure of image components can not only help to guide the construction of the deep networks, but also mine the discrimination of images.

BoW model includes four steps: local feature extracting, code book learning, feature coding and pooling for image classification. In most studies [15–18], feature coding is a key question that involves the feature reconstruction from the structure relationship between local feature and visual words (words structure), which can effectively improve the robustness of feature coding and the

* Corresponding author.
E-mail address: lgf78103@xaut.edu.cn (G. Lin).

http://dx.doi.org/10.1016/j.inffus.2016.12.010
1566-2535/

accuracy of image classification [19,20]. There are two research tendencies for words structure in feature encoding, as follow.

One tendency pays attention to how words structure to effect on coding processing. For example, to decreasing large quantization error between visual words and local feature, soft coding (SC) [15] describes words structure via kernel density estimation and obviously enhances the feature discrimination. In terms of locality and sparsity, local coordinate coding (LCC) [20], local soft coding(LSC) [21] and locality-constrained linear coding (LLC) [17] solve feature sparse representation via words structure constraint.

The other tendency focuses on the accurate description of words structure. For instance, the k-dense neighbors of visual words is found to redefine words structure by the shrinking and expansion algorithm [22] and ameliorate the classification performance. High-order topology is exploited to enhance the robustness of feature coding and boost the accurate rate of classification [23]. There studies have built the model for the neighbor visual words structure in feature reconstruction and obtained the promising achievement.

However, words structure only reveals the principle of feature reconstruction. The distribution structure of images (images structure) that can directly influence the classification performance is less considered for feature coding in the unsupervised way in BoW model. Most studies concentrate on image structure with the different ways in the global feature and involve some methods that are principal component analysis (PCA) [24], independent component analysis (ICA) [25], locality preserving projection (LPP) [26], and canonical correlation analysis (CCA) [27]. From above mention, words structure and images structure respectively effect on image classification, therefore, we mainly explore the both structures fusion mechanism for discovering the discriminative and complementary structure information in feature coding.

To improve the efficacy and discriminability of feature coding, we propose a new local visual feature coding based on heterogeneous structure fusion (LVFC-HSF) algorithm which can explicitly balance the variabilities between words structure and images structure by reformulating the unified feature projection. In the proposed LVFC-HSF algorithm, we jointly formulate words structure and images structure in one optimization framework. The objective function of LVFC-HSF unites two structures in a closed form which can be optimized alternately via linear programming and generalized eigenvector. The LVFC-HSF solution provides not only the optimally local feature projection but also the weight coefficients that encode the relative importance and relevance between two kinds of structures.

The rest of the paper is organized as follows. Section 2 briefly introduces the background of feature coding in BoW model. Section 3 describes the formulation and optimization of the proposed LVFC-HSF algorithm. Section 4 details a comprehensive performance evaluation of LVFC-HSF in four image database, where LVFC-HSF is thoroughly examined and compared with different coding methods as well as the state-of-the-art algorithms in each database. Finally, the conclusions are summarized in Section 5.

2. Related works

The studies on feature coding with words structure constrain can be divided into two kinds of methods in terms of the structure relation formation. The first one tries to find the better description of local structure for the robustness of feature coding. The spatial pyramid matching (SPM) [28] model separates the spatial region of image to represent the different scales structure by vector concatenation for the effective image recognition. However, vector concatenation can not quantify the related structure. To further describe the structure relationship, some studies consider not only spatial structure by vector concatenation, but also words structure via distance matric (for example, Euclidean distance) [29,30]. These methods usually obtain the several nearest neighbors(local structure), which could include the weaker structure relation because of the long distance and the structure confusion owing to the orientation loss. Therefore, k-dense neighbors search enhances the structure relationship of neighbors [22] and the orientation distribution restrains the noise in large image variations [31] for good performance. In addition, high-order topology is further used to explain the local structure measurement for the better performance [23].

The second methods focus on the different structures combination for feature coding. The density of structure considers not only the relationship between local feature and visual words, but also the pairwise relations among visual words [22,23]. The structure combinations of various distance metric and orientation restrictions can better adapt to the different database [32]. These methods demonstrate that the different structures combination can improve image classification performance, and mostly involve homogenous structure, which intrinsically is the local structure of visual feature in despite of it's different metric or various constraint (spatial or oriented limitations). However, heterogeneous structure is also an important fact to enhance the discriminability of the feature [33]. In our research [33–37], the distribution structure between the global multi-feature or the correlation structure among the global multi-feature sets is respectively or jointly optimized for feature fusion. In addition, high-order structure based on hypergraph is constructed to describe the manifold of data for multi-modality (multi-view,multi-feature) learning, which shows the promising results [38–40]. In this paper, the difference is that words structure of the local feature and images structure of the samples in single feature is fused to approximate the intrinsically discriminative structure of image for feature encoding. Moreover, LVFC-HSF exploits the nonlinear relationship of both structures to cognise the operation mechanism of heterogeneous structure in feature encoding. Fig. 1 shows the main idea of this paper.

3. Local visual feature coding based on heterogeneous structure fusion (LVFC-HSF)

This section firstly introduces high-order topology [23] to present words structure, and then, uses the statistics characteristic of words to describe images structure. Finally, this section explains how to fuse words structure and images structure in this algorithm, and utilizes the alternate optimization to learn not only the projection matrix for feature coding but also the structure fusion coefficient vector for mechanism revealing.

3.1. Notation and problem statement

M local visual features $\mathrm{x}_1, \ldots, \mathrm{x}_M \in \mathbb{R}^D$ come from N images, and then M_1 visual words $\mathrm{c}_1, \ldots, \mathrm{c}_{M_1} \in \mathbb{R}^D$ is learned from M local visual features by k-mean algorithm. $\mathcal{X}_i = \{\mathrm{x}_{i1}, \ldots, \mathrm{x}_{in_i}\}$ represent the local feature set in image i. Therefore, $\mathcal{F} = \{\mathcal{X}_1, \ldots, \mathcal{X}_N\} = \{\mathrm{x}_1, \ldots, \mathrm{x}_M\}$ is all local features set in N training images. To capture the discriminative structure of images, we find the projection matrix $A \in \mathbb{R}^{D\times d}$ for transforming the local feature to the low-dimension space with words structure and images structure constraints. However, words structure is the relationship between visual words, while images structure is the distribution among images. They are heterogeneous owing to the different description object (one is the visual words, the other is the images). To balance and complement these structures, we simultaneously need explore the coefficient vector ω of these structures fusion for describing the nonlinear relationship of heterogeneous structure. Therefore, we can construct the unified optimization function $J(A, \omega)$ in Section 3.4 and define the