Information Sciences 460–461 (2018) 115–127

Contents lists available at ScienceDirect

Information Sciences

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

Birds of a feather flock together: Visual representation with scale and class consistency

Chunjie Zhang[a,b], Chenghua Li[c], Dongyuan Lu[d,*], Jian Cheng[a,b,c,e], Qi Tian[f]

[a] *Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[b] *University of Chinese Academy of Sciences, Beijing 100080, China*
[c] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing, China*
[d] *School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China*
[e] *Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China*
[f] *Department of Computer Sciences, University of Texas at San Antonio, TX 78249, USA*

ARTICLE INFO

Article history:
Received 7 November 2016
Revised 7 January 2018
Accepted 22 May 2018
Available online 22 May 2018

Keywords:
Scale consistency
Discriminative sparse coding
Class consistency
Image classification

ABSTRACT

There are three problems with a local-feature based representation scheme. First, local regions are often densely extracted or determined through detection without considering the scales of local regions. Second, local features are encoded separately, leaving the relationship among them unconsidered. Third, local features are simply encoded without considering the class information. To solve these problems, in this paper, we propose a scale and class consistent local-feature encoding method for image representation, which is achieved through the dense extraction of local features in different scale spaces, and the subsequent learning of the encoding parameters. In addition, instead of encoding each local feature independently, we jointly optimize the encoding parameters of the local features. Moreover, we also impose class consistency during the local-feature encoding process. We test the discriminative power of image representations on image classification tasks. Experiments on several public image datasets demonstrate that the proposed method achieves a superior performance compared with many other local-feature based methods.

1. Introduction

Local features have recently demonstrated their effectiveness with regard to image classification. Local features are often used in a bag-of-visual-words (BoW) manner [41]. However, the quantization loss is heavy when local features are assigned to the nearest visual word. To reduce the amount of quantization loss, researchers have proposed various soft-assignment based methods [7,9,20,25,29,44], of which a sparse coding based technique is widely used. Sparse coding attempts to minimize the summed reconstruction error with a sparsity constraint. Max pooling is then used to extract the image representations.

To cope with image variances, researchers often densely extract local features with multiple scales [12,14]. However, the scale information is simply discarded. Object detection based methods [23,26] help to alleviate this problem. However, they

[*] Corresponding author at: School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China.

E-mail addresses: chunjie.zhang@ia.ac.cn (C. Zhang), chenghua.li@ia.ac.cn (C. Li), ludy@uibe.edu.cn (D. Lu), jcheng@nlpr.ia.ac.cn (J. Cheng), qitian@cs.utsa.edu (Q. Tian).

https://doi.org/10.1016/j.ins.2018.05.048
0020-0255/© 2018 Elsevier Inc. All rights reserved.

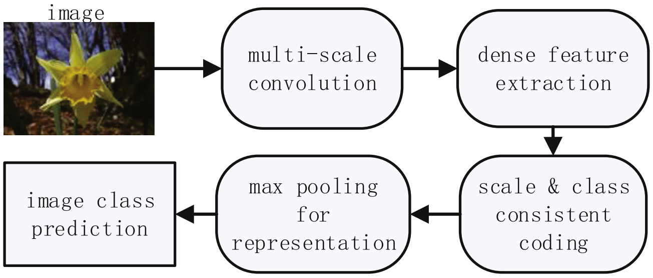

Fig. 1. Flowchart of the proposed scale and class consistent local feature encoding method for image representation and classification.

rely heavily on the accuracy of the detection results. In addition, they require more labeled samples and computational power. To avoid the explicit detection of objects, researchers have also attempted to use salience-based methods. However, the aim of a salience measurement is inconsistent with image classification. In fact, the local feature extraction process has plenty of information that can be used. For example, we often densely extract local features with multiple scales. However, the scale information, which can be used to measure the relative sizes of local regions, is often ignored. If two images of the same class have different sizes, the scale information will help represent them discriminatively. If we can make use of the scale information along with visual similarity in a unified framework, we will be able to represent images more effectively.

Thousands of local features are often extracted from a single image. To encode these features, researchers often iteratively encode each local feature [25], or use an online feature encoding technique [12]. However, local features are inherently correlated. To jointly encode local features, the nearest neighbor information is used [7,44]. In addition, the spatial relationships of local features have also been widely explored [14,21,34]. However, visually similar features may belong to different classes. When considering only the visual similarities of local features, it might not be possible to fully explore the useful information of the local features. The class information should therefore also be used to boost the classification performance [18].

To make full use of local feature information for image representation and classification, in this paper, we propose a novel discriminative scale and class consistent local feature encoding technique for image representation. Instead of only using the visual information of the local features, we also consider their scale and class information, which is achieved by first mapping the original image to multi-scale spaces through a convolution, and then densely extracting the local features. To explore the correlations of the local features, we jointly optimize for the encoding parameters with scale consistency. In addition, the class information of the local features is also combined to encode the local features. In this manner, we are able to obtain more effective image representations than other local feature-encoding based methods. Image classification experiments on several public image datasets have proven the effectiveness of the proposed method. Fig. 1 provides a flowchart of the proposed method.

There are three main contributions of this study:

- First, we use the scale information of the local features along with visual similarities to encode the local features.
- Second, the class information of the local features is also used to boost the discriminative power of the image representations.
- Third, we jointly encode the local features with a sparsity constraint to achieve a superior image classification performance over other baseline methods.

The rest of this paper is organized as follows. Related works are described in Section 2. The details of the proposed scale and class consistent visual representation method are provided in Section 3. Image classification experiments conducted on several public image datasets are described in Section 4. Finally, some concluding remarks are given in Section 5.

2. Related work

The BoW model has been widely used for image representations. It first extracts the local features and then encodes them based on a nearest neighbor assignment. To reduce the quantization loss, the use of a soft-encoding strategy has become popular [7,9,20,22,25,26,44,50]. Gemert et. al [9] used a kernel trick by softly assigning a number of visual words instead of only one. Zhang et al. [35] made use of the sparse coding technique for image classification. Wang et al. [25] constrained the sparse coding with locality to improve the performance. Gao et al. [7] proposed the use of Laplacian sparse coding to ensure that visually similar local features are encoded using similar parameters. To avoid inconsistency between sparse coding and max pooling for an image representation, Zhang et al. [44] proposed the use of non-negative sparse coding. To encode local features more finely, Sanchez et al. [20] proposed the Fisher vector technique.

The extraction of local features with multiple scales has been widely adopted. Han et al. [11] proposed a novel method for combining multiple cues for discriminative descriptor learning with an improved level of performance. Lazebnik et al. [14] extracted SIFT features with multiple scales. Kulkarni and Li [12] used an affine transformation and then extracted the local features. Wang et al. [26] used the feature context with pre-defined scales for image classification. However, the computational cost is high, although the combination of object detection and classification [23] can alleviate this problem