Neurocomputing 000 (2017) 1–9

Contents lists available at ScienceDirect

Neurocomputing

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

Categorizing scenes by exploring scene part information without constructing explicit models

Shuang Bai*, Huadong Tang

School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Article history:
Received 22 February 2016
Revised 20 November 2017
Accepted 1 December 2017
Available online xxx

Communicated by Prof. Zidong Wang

Keywords:
Scene categorization
Scene part
Cascading framework
Random forest
Support vector machine

ABSTRACT

Approaches based on scene parts are deemed to be one of the main streams for scene categorization. In previous methods, before one can utilize scene parts, models need to be constructed for them first. The quality of part models has a great influence on the final results. However, building high-quality scene part models is still an open question. To perform scene categorization based on parts effectively, in this paper we propose to explore scene part information without constructing explicit models for them. For this purpose, a cascading framework is used, at each of whose stages we aim to process image patches potentially corresponding to scene parts from different perspectives. Specifically, the first stage of the framework uses the selective search algorithm to extract possible object patches from images and represents obtained patches based on convolutional neural networks. Then, spectral clustering and linear support vector machines are adopted to select representative visual patterns for images in the second stage. In the third stage, random forest and multi-class support vector machines are combined to mine and classify image features for determining the categories of the images. Through using the cascading framework, we can explore scene part information step by step without needing to construct explicit models for them. Finally, extensive experiments are conducted to evaluate the proposed method on three well-known benchmark scene datasets, i.e. MIT Indoor 67, SUN397 and Places. Experiment results demonstrated the effectiveness of the proposed method.

1. Introduction

Scene categorization means to classify scene images into different semantic classes based on their contents. As one of the main problems in computer vision, it has long been deemed to be a challenging task due to huge intra-class variations and inter-class ambiguities of scene images. Scene categorization has a wide range of applications such as image browsing, retrieval and understanding, where extensive knowledge from scene categories can provide much information beyond the images themselves [1].

In order to perform scene categorization effectively, various approaches have been proposed. Recently, methods based on scene parts have become one of the main streams in this field [2–4]. In fact, part based methods were originally introduced for solving the object recognition problems [5,6]. With the realization that scene categories are closely related to objects (scene parts) appearing in them, Quattoni and Torralba propose to take manually segmented regions of scene images as parts for scene categorization [3]. Following their work, many techniques are presented to explore parts in scene images [2,4].

In previous methods, to categorize scene images by utilizing parts, part models need to be constructed first. Generally speaking, there are two main ways of constructing part models. The first way is unsupervised or weakly supervised, in which scene parts are utilized without being assigned any semantic meanings. In such methods, the learning of part models is usually initialized randomly or with simple heuristics [7,8]. Then, the models are obtained based on iterative mining procedures [7,9]. Unsupervised or weakly supervised part learning methods are sensitive to the initialization [10]. Without good initialization, their performances will be affected.

The other way of constructing part models is supervised [11,12]. In supervised methods scene parts need to be defined manually and trained with clearly annotated samples. Therefore, considerable amount of human efforts are required for providing detailed annotations. Such methods are labour intensive to obtain accurate data labelling and intrinsically ambiguous to define scene parts [13].

So far, building appropriate parts for scene categorization is still an open question. To categorize scene images effectively based on

* Corresponding author.
E-mail address: shuangb@bjtu.edu.cn (S. Bai).

https://doi.org/10.1016/j.neucom.2017.12.003
0925-2312/

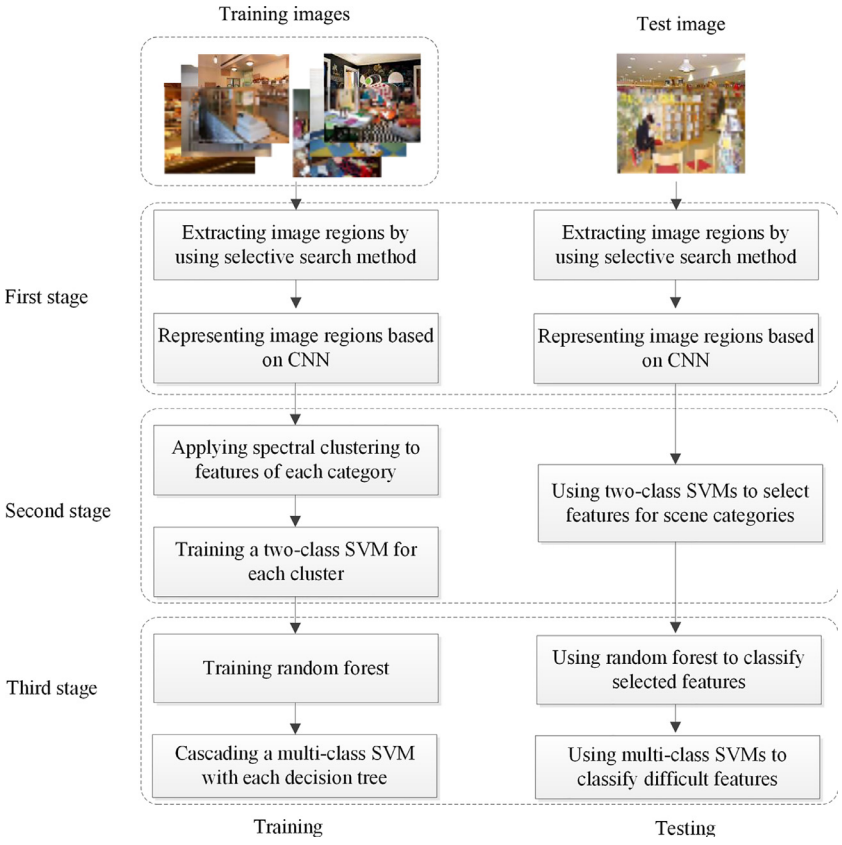

Fig. 1. Framwork of the proposed method.

parts, in this paper we propose to explore scene part information without constructing explicit models for them. For this purpose, we adopt a cascading framework in which we can distinguish image patches step by step for determining the categories of images. At each stage of the framework, rather than constructing scene part models, we only focus on selecting and classifying visual patterns that are useful or difficult for scene categorization.

Additionally, previous methods on using parts for scene categorization mainly utilize low-level features such as HOG [14] or SIFT [8] for constructing part models. Because low-level features are not powerful enough, performances of methods employing them are limited. On the other hand, deep learning, which refers to algorithms that automatically learn hierarchical features from raw data based on deep architectures, has had great achievements in a number of applications, e.g. speech recognition [15], computer vision [16,17] and natural language processing [18] etc. Particularly, convolutional neural network, which uses convolution operators in each layer to map inputs to new representations via a bank of filters [19], is highly suitable for modelling images [20]. To obtain better results, in this paper we adopt convolutional neural network to extract features from images for exploring scene part information for scene categorization.

The framework of the proposed method is given in Fig. 1, which consists of a training procedure and a testing procedure. In the first stage of training, we first extract possible object patches from training images by using the selective search algorithm [21] and represent each image patch based on convolutional neural network [17]. Then, in the second stage we aim to construct models that are able to recognize representative visual patterns for images. To do so, we apply spectral clustering to features of each scene category to mine representative visual patterns for it. After that, for each cluster a two-class support vector machine is trained, where features in the cluster are used as positive samples, while features from other categories are used as negative samples.

In the third stage, we aim to mine and classify features potentially corresponding to scene parts by combining random forest and multi-class support vector machines. We first use image features to train a random forest. Then, we pass all the training samples through each decision tree of the random forest respectively to evaluate its leaf nodes. For a decision tree, training samples passed to its leaf nodes whose correct rate is lower than a given threshold are collected to train a multi-class support vector machine. As a result, each decision tree in the random forest has a support vector machine combined with it. In the testing procedure, given an input image, it is first processed as in the training stage for extracting image patches and creating features. Then, obtained features are classified by the support vector machines corresponding to feature clusters for selecting representative ones, which are fed to the third stage. In the third stage, each selected feature is classified by the decision trees in the random forest. For a decision tree, if the given feature is passed to its leaf node with a correct rate higher than the given threshold, the predication of the leaf node is taken as the predication of the decision tree, otherwise the multi-class support vector machine combined with the decision tree is used to give a new predication to replace the one given by the decision tree. Finally, the category of the image is determined by the votes of the decision trees in the random forest and multi-class support vector machines combined with them.

The rest of the paper is organized as follows. Section 2 briefly reviews relevant work on scene categorization. Section 3 introduces the image patch extraction and representation procedure of the proposed cascading framework. Section 4 describes the representative feature selection method. In Section 5 we detail how selected image features are classified to determine categories of

Please cite this article as: S. Bai, H. Tang, Categorizing scenes by exploring scene part information without constructing explicit models, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.12.003