Pattern Recognition 84 (2018) 165–181

Contents lists available at ScienceDirect

Pattern Recognition

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

Parametric and nonparametric context models: A unified approach to scene parsing

Parvaneh Aliniya, Parvin Razzaghi*

Artificial Intelligence Lab, Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran

ARTICLE INFO

Article history:
Received 13 September 2017
Revised 18 June 2018
Accepted 8 July 2018
Available online 17 July 2018

Keywords:
Scene parsing
Parametric
Nonparametric
Context model
Co-occurrence graph
Unified framework

ABSTRACT

In this paper a new nonparametric scene parsing approach is proposed which has three steps: image retrieval, label transferring and label gathering. In our approach, to incorporate the contextual knowledge in scene parsing, we propose to integrate both parametric and nonparametric context models into a unified framework. We adopt a co-occurrence graph to be our parametric context model to learn the co-occurrence frequency of objects. To consider different preferences of the co-occurring of one object with the other objects, the concept of co-occurring priority is introduced in this paper for the first time. Next, by using the learned co-occurrence graph and the context knowledge of the set of retrieved images, we propose new ways to incorporate contextual information in all three steps of nonparametric scene parsing approach. To evaluate our proposed approach, it is applied on MSRC-21 and SiftFlow datasets. The results show that our approach outperforms its competitors.

1. Introduction

Scene parsing has received much interest in the recent decades. The aim is to provide a semantic label for each pixel in an image using a predefined set of labels. In other words, in scene parsing, each image is segmented or parsed into regions associated with semantic categories. Traditional parametric approaches of scene parsing learn an object model for each object category. Learning models and their corresponding parameters are estimated during the training stage. Assume that we want to add new object categories to the existing system. To do this, we need to learn a new model for new object categories which is often tedious and time-consuming. In contrast, in nonparametric approaches, instead of learning sophisticated models for each object category, the knowledge from the labeled training images is transferred to the unlabeled image. Typical nonparametric approaches in scene parsing have three main steps. In the first step, we retrieve a small subset of training images which are visually similar to the query image. In the second step, labels from the retrieved training images are transferred to the query image. Up to this point, different labels may be assigned to each pixel. In the third step, to aggregate labels, Conditional Random Field (CRF) framework is used. In summary, these methods, match a query image to the existing set of annotated images. Then, labels from the annotated images are transferred to the query image. It should be noted that the imperfection in each step leads to a decrease in the final scene parser performance. The most important advantages of nonparametric approaches are independent from the dataset and the number of object categories. Also, these approaches do not need learn model parameters for each dataset once again. Hence, in this paper, we propose a nonparametric approach to scene parsing.

Most methods in scene parsing rely on visual appearance information of local patches to divide images into semantic regions. Still, we need to consider high-level semantic information of scenes to achieve the desirable results. The incorporation of high-level semantic information such as object level and context level information is especially effective in scene parsing. The context knowledge considers semantic relationships among the different object categories (e.g., 'grass' often co-occur with 'trees' or the 'sky' is above the 'ground') which enable machines to perceive an image in a similar manner to a person. However, the performance benefit of context models has been limited because most of the previous methods modeled the contextual information in a parametric or nonparametric framework. Parametric context models [4,5] focus on fixed models like, pairwise relationships between object [5] or a context model for each object [4]. As regards, an object may be presented in different scenes (e.g., a cat in indoor or outdoor scenes), hence the fixed context may lead to incorrect labeling. Also, pairwise co-occurrence models may result in missing some sort of contextual information, because they do not consider the spatial layout of the objects in a scene. However, in nonparametric context models, context knowledge is extracted from the set of retrieved images. Hence, the spatial layout of objects in a specific scene can be considered. In nonparametric context mod-

* Corresponding author.
E-mail addresses: palinia@iasbs.ac.ir (P. Aliniya), p.razzaghi@iasbs.ac.ir (P. Razzaghi).

https://doi.org/10.1016/j.patcog.2018.07.013
0031-3203/

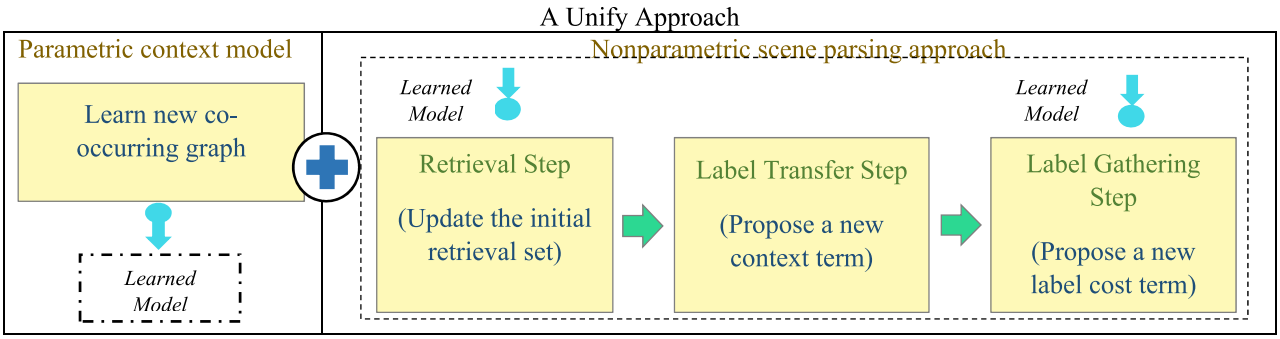

Fig. 1. The overall scheme of the proposed approach. Boxes show the contributions of this paper.

els, the performance of the methods depends on the quality of the retrieved images. Therefore any incorrect retrieved image may decrease the performance of the method. Beyond the importance of context knowledge in improving the performance of scene parsing, there is lack of modeling (e.g., parametric or nonparametric point of views) which leads in practice to imperfect utilization of the contextual knowledge.

As a result, we find out that parametric and nonparametric context models are complementary to each other for performance boosting. Hence, we propose a new nonparametric scene parsing approach which effectively utilizes the contextual knowledge of the parametric and nonparametric context models.

Moreover, the human visual system is another intuition of the proposed approach. Cognitive studies have provided important insights about how the brain uses the contextually associated information in scene understanding. By investigating the role of contextual information in the human visual system, it is found that the typical arrangements of objects in the environment, have an effective role in scene understanding [6]. Object arrangements in a scene are represented in context frames. Context frames are contextual structures that integrate information about the identity of the objects that are most likely to appear in a specific scene with information about their relationships with the other objects. It provides the expectations that facilitate the perception of the other scenes by human that can be represented in the same context. Moreover, in [7,8] it is mentioned that human, by seeing a new scene, retrieve similar scenes from episodic memory which depends upon contextual information or cues. Episodic memory is a neurocognitive (brain/mind) system which enables human beings to remember the past experiences. In the process of forming an episodic memory, the events or scenes are taken into the long-term memory. As a result, human brain benefits from context frames and episodic memory simultaneously. Hence, inspired by the human visual system, in our proposed approach, we combine context frames as parametric context model and episodic memory as nonparametric context model to incorporate contextual information in scene parsing.

We adopt a new proposed co-occurrence graph to be our parametric context model to learn the frequency and priority of co-occurrence as well as geometrical properties of objects. The co-occurrence priority considers the different preferences of the co-occurring of one object with the other objects which is proposed in this paper for the first time. Next, in the test phase, by using the learned co-occurrence graph and the context knowledge of the set of retrieved images, we propose new ways to incorporate contextual information in all three steps of nonparametric scene parsing approach. As mentioned, we have three steps: retrieving step, label transferring and label aggregating step. Each step has significantly impact on the final scene parsing result. In the proposed approach, we adopt the context knowledge in each step to increase the reliability of each step result. To do this, the context knowledge from the retrieved images and the proposed co-occurrence graph are utilized in the following ways: 1) In the retrieval step, after the initial retrieving, the retrieval set is updated using the contextual information of the retrieved images and the learned co-occurring graph. The retrieval set is updated such that all the retrieved images have consistent contextual association. 2) In the label transferring step; a new context term is added to the label transferring energy function. The proposed context term is designed such that it transfers the overall spatial layout of semantic labels in the training image to the query image. Our proposed context term is different from the similar ones. For example, [1] encodes the relative center positions of each semantic class in image with respect to other classes. Since the center position of each semantic class in the query image is affected by the error, consequently, the context term is not reliable, 3) In the aggregating step, by combining the context knowledge of the learned graph and the retrieved images, a new label cost term for conditional random filed (CRF) is proposed. In other words, in our approach, the context knowledge in all steps is incorporated. The overall scheme of the proposed approach is shown in Fig. 1. The most important novelty of our approach is the boxes which are highlighted in Fig. 1. Also, the detailed scheme of the proposed approach is shown in Fig. 2.

To sum up, the main contributions of this paper are given as follows: 1) combining parametric and nonparametric contexts models into a unified approach, 2) proposing a new co-occurrence graph which not only considers the frequency of co-occurrence of objects but also considers their co-occurring priority (see Fig. 2-a), 3) proposing an approach which updates the set of retrieved images by utilizing the contextual information of the set of initial retrieved images and learned co-occurrence graph (see Fig. 2-b), 4) proposing a new context term which is added to the energy function in the label transferring step (see Fig. 2-c), 5) proposing a new label cost term which considers the contextual information in the label gathering step (see Fig. 2-d).

The rest of the paper is organized as follows. The related work is presented in Section 2. In Section 3, the proposed approach is given. Section 4 shows the results of applying our proposed approach to the best well known MSRC-21and SiftFlow datasets. Concluding remarks are given in Section 5.

2. Related work

Early studies have shown that context is a rich source of information. Many recent works leverage contextual information for enhanced recognition and localization of objects in images. Various sources of context have been studied, such as the global scene layout, different kinds of objects and scene interaction as well as local features. So far, for most of the computer vision tasks, many approaches based on using contextual information have been introduced.