Pattern Recognition Letters 42 (2014) 56–64

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A new fast approach to nonparametric scene parsing ☆

Parvin Razzaghi [a,*], Shadrokh Samavi [a,b]

[a] Electrical and Computer Engineering Department, Isfahan University of Technology, Isfahan, Iran
[b] Electrical and Computer Engineering Department, McMaster University, Hamilton, Canada

ARTICLE INFO

Article history:
Received 10 July 2013
Available online 24 January 2014

Keywords:
Nonparametric scene parsing
Label transferring
High level information
Graphical model

ABSTRACT

Scene parsing is a challenging research area in computer vision. It provides a semantic label for each pixel in image. Most scene parsing approaches are parametric based which need a model that is acquired through a learning stage. In this paper, a new nonparametric approach to scene parsing is proposed which does not require a learning stage. All introduced nonparametric approaches are based on patch correspondence. Our proposed method does not require explicit patch matching which makes it fast and effective. The proposed approach has two parts. In the first part, a new generative approach to transfer semantic labels from a training image to an unlabelled test image is proposed. To do this, a graphical model is constructed over regions of both the training and test images. Then, based on the proposed graphical model, a quadratic convex function is defined on likelihood probability of each region. Cost function is defined such that contextual information and object-level information are both considered. In the second part of our approach, by using the proposed method of transfer knowledge, a new nonparametric scene parsing approach is given. To evaluate the proposed approach, it is applied on the MSRC-21, Stanford background, LMO, and SUN datasets. The obtained results show that our approach outperforms comparable state-of-the-art nonparametric approaches.

1. Introduction

Scene parsing has received much interest in recent decades. The aim is to provide a semantic label for each pixel in image using a predefined set of labels. Up to know, many approaches in scene parsing are introduced. These approaches can be divided into two groups: parametric approaches or nonparametric approaches.

Parametric approaches are learning based methods. Hence, learning models and their corresponding parameters are estimated during the training stage. These approaches are dependent to training samples. Model parameters should be updated when a new sample is added to the system. Also, for a new dataset, it is almost needed to learn models once again. There are many approaches in this field. Most of these approaches use Conditional Random Field (CRF) over regions which in standard form have two terms: data term which considers appearance information of each region and a smoothness term which encourages similar neighboring regions to have the same labels. In standard CRF, high order dependencies between nodes of the graph (or regions in image) are not considered. Hence, many approaches are introduced to incorporate high order information by encouraging some set of pixels to have the same label [10,17,18,21], incorporating object detection results [1,7,20,40], or incorporating context information [6,22,28,35,8].

In nonparametric approaches, instead of learning sophisticated models, knowledge from the labeled training samples is transferred to the unlabelled image. These approaches not only have the competitive performance with the state-of-art parametric approaches but also have several key advantages. The most important advantages of nonparametric approaches are the independence from the dataset and the independence from the number of object categories. Also, these approached do not need to learn model parameters.

Typical nonparametric approaches in scene parsing have three main steps. In the first step, for each test image, small set of similar images from the training set are retrieved. In the second step, correspondence map between regions of images in the retrieval set and regions of the test image is obtained. Then, from the retrieved set, labels are transferred to the test image via dense region mapping. Up to this point, different labels may be assigned to each pixel. In the third step, to aggregate labels, Markov Random Field (MRF) framework is used. Liu et al. [24] use SIFT flow [25] to find pixel mapping. They extend SIFT flow algorithm to improve the performance. Their approach finds a smooth pixel mapping between test image and images of the retrieved set. Finally, to integrate multiple semantic labels of each pixel, MRF is established. Zhang et al. [41] use KNN–MRF matching schema to find a correspondence map. To do this, for each test image, some small set of

☆ This paper has been recommended for acceptance by R. Davies.

* Corresponding author. Tel.: +98 311 391 2450; fax: +98 311 391 2451.
E-mail address: p.razzaghi@ec.iut.ac.ir (P. Razzaghi).

http://dx.doi.org/10.1016/j.patrec.2014.01.003
0167-8655/

similar training images are retrieved. The test image and all training images are segmented into superpixels. Then, for each superpixel in the test image, K similar superpixels from the training images are achieved. To find final correspondences of each superpixel, MRF is used. Next, to reduce the incorrect correspondences, a set of classifiers for each category is learned. Finally, by using the output confidence value of classifiers and smoothness term, another MRF to label each pixel is established. The most important bottleneck of these approaches is the time consuming step of finding correspondences map of pixels which is done during the test phase. To overcome this difficulty, Gould and Zhang [13] construct a graph of patch correspondences by using all images of the dataset during the training phase. To do this, PatchMatch algorithm [3] is modified so that for each patch there are K nearest patches. The modified algorithm is called PatchMatchGraph. One shortcoming of their approach is that PatchMatchGraph is constructed over all images (training and test sets) of dataset in the training phase. If PatchMatchGraph were only constructed for the training images, then in the test phase, pairwise patch correspondence would have been needed. Recent nonparametric approaches try to employ contextual information to obtain more accurate results. Myeong et al. [29] introduce a new data driven approach in which contextual relationships between all pair of regions in the labeled image are transferred to the unlabelled image by applying link analysis technique [27]. Myeong and Lee [30] add a new term to the CRF. Data and smoothness terms are computed based on conventional nonparametric approach [36,37]. In a new term, high order semantic relations between objects are transferred from labeled images to the unlabelled image. In their approach, third order semantic relation between region triplets is considered and is referred as semantic tensor. To transfer this high order information, an objective function is defined over semantic tensors.

In the present paper, we propose a new fast nonparametric approach to scene parsing which does not require pairwise patch correspondences. Also, in our nonparametric approach, contextual information and object-level information are jointly considered. Whereas, in most approaches based on patch correspondence, only local patch information and in some cases, contextual information are used and object level information is not considered. The proposed approach has two parts. In the first part, a generative approach to transfer semantic labels from one training image to an unlabelled test image is proposed. To do this, each region in test image is assigned to a region in training image. To estimate the belonging degree of each region of test image to regions of training image, a graphical model is constructed over regions of both training and test images. Next, by using the proposed graphical model, a quadratic convex function is defined over regions likelihood. Then, by optimizing quadratic convex function, likelihood probability of each region of test image is estimated. It should be noted that contextual information and object-level information is encoded in cost function definition. Our proposed method to transfer knowledge is designed so that it does not require explicit patch matching. Hence, it is fast and effective. In the second part of our approach, it is shown how the proposed generative method of transfer knowledge is used to multi class pixel labeling. In this case, for a test image, a small set of similar images are retrieved. Then for each pair of retrieved training image and test image, knowledge is transferred using the proposed generative method. Finally, MRF framework is used to aggregate knowledge and assign semantic label to each pixel.

The main contributions of this paper are as follows: (1) Providing a new generative approach to transfer semantic knowledge from one training image to test image. (2) Defining cost function over regions likelihood such that object level information and contextual information are jointly included. (3) Propose a new fast nonparametric approach to scene parsing which does not require patch matching in the training and test phases.

The rest of the paper is organized as follows. The transfer of knowledge from one training image to a test image is presented in Section 2. In Section 3, a new nonparametric approach to scene parsing is given. Section 4 shows the results of applying our proposed approach to the best well known MSRC-21, Stanford background, LMO and SUN datasets. Concluding remarks are given in Section 5.

2. Transfer semantic labels

In this section, a generative approach to transfer semantic label from one labeled image to an unlabelled image is proposed. In the following, at first, the problem formulation is given. Then, graphical model of our approach and how to model the likelihood probability of each region are discussed.

2.1. Problem formulation

Given I^t as a training image in which each pixel is assigned one semantic label $l \in \{1, 2, \ldots, B\}$. Let B to be the number of class labels in image I^t. The regions of image I^t are represented by $R^t = \{r_i^t\}_{i=1}^M$ and semantic label of each region is shown by $l(r_i^t)$. Let M to be the number of regions in image I^t. It should be noted that $M > B$ since it is possible that there are two distinct regions with the same class label. Let I denotes a test image which is unlabelled. At first, test image is segmented into some regions which are represented by $R = \{r_i\}_{i=1}^N$. Let N denote the number of regions in the test image. Our goal is to assign a semantic label to each region in image I. To do so, we assume function $g : R \rightarrow R^t$ assign each region in image I to one region in image I^t. Then, the label of the corresponding region in image I^t is transferred to a region in image I.

We use a generative approach to determine function $g(\cdot)$. It should be noted that, groups of regions in test image I can be assigned to one region in training image I^t. Conceptually, regions in test image I are grouped to form one semantically meaningful unit.

To obtain $g(\cdot)$, a generative approach is proposed. Hence, we have:

$$g^t = \arg\max_g p(g|I; I^t) = \arg\max_g \prod_i p(g_i|r_i; I^t)$$

$$\propto \prod_i \arg\max_{g_i \in \{r_j^t\}_{j=1}^M} p(r_i|g_i; I^t) p(g_i; I^t)$$

where g_i is the short form for $g(r_i)$ which it denotes the corresponding region of r_i in image I^t. It is noted that $g_i \in \{r_j^t\}_{j=1}^M$. In generative approaches, likelihood probability $p(r_i|g_i; I^t)$ and prior probability $p(g_i; I^t)$ are directly modeled. In our approach prior probability $p(g_i; I^t)$ is assumed to be uniform. Also, to model $p(r_i|g_i; I^t)$, a graphical model is constructed. In the following, how to model the likelihood probability is explained. Likelihood probability modeling is motivated from [16] in which a nonparametric approach to interactive segmentation is introduced.

2.2. Graphical model

In our model, a graphical model is constructed in which training image regions (V^t) and test image regions (V^q) are as nodes of a graph. Training image is labeled manually, hence each connected component with a same semantic label is considered as one region. However, to obtain regions in the test image, it is segmented into regions using the approach of [2]. It identifies boundaries well and the number of produced regions is relatively low. Also regions are uniform in terms of brightness, color or texture. However, in this step, other image segmentation algorithms can be used too.