Pattern Recognition Letters 70 (2016) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Label transfer via sparse representation☆

Taeg-Hyun An*, Ki-Sang Hong

Department of Electrical Engineering, POSTECH, Namgu Pohang, Republic of Korea

ARTICLE INFO

Article history:
Received 21 January 2015
Available online 27 November 2015

Keywords:
Scene parsing
Sparse representation
Boosting

ABSTRACT

In this paper, we present a simple and effective approach to the image parsing (or labeling image regions) problem. Inspired by sparse representation techniques for super-resolution, we convert the image parsing problem into a superpixel-wise sparse representation problem with coupled dictionaries related to features and likelihoods. This algorithm works by image-level classification with global image descriptors, followed by sparse representation based likelihood estimation with local features. Finally, Markov random field (MRF) optimization is applied to incorporate neighborhood context. Experimental results on the SIFTflow dataset support the use of our approach for solving the task of image parsing. The advantage of the proposed algorithm is that it can estimate likelihoods from a small set of bases (dictionary) whereas recent nonparametric scene parsing algorithms need features and labels of whole datasets to compute likelihoods. To our knowledge, this is the first approach that utilizes sparse representation to superpixel-based image parsing.

1. Introduction

Scene understanding at many different levels has been researched over the past few decades. At the image-level, we may consider the general category of the scene, e.g., coast, forest, street with global image descriptors [1,13]. Scene parsing is a pixel-level problem which assigns a semantic label (e.g., sky, cars, buildings) to every pixel in an image. Although it is a challenging problem, research in this field has been active due to its various applications, e.g., image editing, autonomous robots, surveillance system. Researchers have proposed a lot of approaches for scene parsing. Due to the complicated nature of the problem, these approaches have different choice of features to describe, regions to localize, relationships to incorporate context and optimization techniques to solve. Typically, Conditional Random Fields (CRFs) based approaches are successfully used to solve the scene parsing problem [7,9] and even combined with object detection algorithms [10,22]. Deep learning techniques are also used to obtain feature maps with a trained neural network [4]. Recently, nonparametric approaches have been researched as the sizes of datasets and the number of labels have increased [12,15]. These are data-driven methods that utilize nearest neighbor (NN) search techniques. With the simplicity and effectivity of nonparametric approaches, researchers have proposed advanced techniques.

[12] used "SIFT flow" which estimates dense correspondences between two images. Given a query image, the algorithm first finds similar images within annotated dataset, i.e. image retrieval, and estimates "SIFT flow" from the query image to each similar image. Then, they "transfer" labels from the dataset to the query image by using estimated flow. [15] proposed the "SuperParsing" algorithm which analyzes superpixels instead of pixels since pixelwise-correspondence estimation is complex and computationally expensive. Each superpixel is labeled by using NN superpixels within retrieved images, then using MRF inference to incorporate neighborhood context. More recently, they extended the SuperParsing algorithm with per-exemplar detection [16]. [14] proposed a nonparametric based approach that uses a locally-adaptive NN technique and refines the retrieval set by considering spatial pyramids of predicted labels. [20] proposed nonparametric scene parsing algorithm with attention to rare classes; this algorithm achieved state-of-the-art performance.

Example-based super-resolution algorithms have similar nature to those of nonparametric scene parsing algorithms. Both start with features of an input query image (e.g.: low-resolution image patches for super-resolution, local features of superpixels for nonparametric scene parsing) and find similar (e.g.: NNs) features from datasets. Features from datasets have their coupled parts (e.g.: high-resolution image patches for super-resolution, ground truth label information of superpixels for nonparametric scene parsing) and both use the information of coupled parts with similarities found in feature domain. Recently, sparse reconstruction based super-resolution algorithms have been proposed instead of using whole dataset. For example, [21] proposed a joint dictionary learning model that uses concatenated high resolution (HR)/low resolution(LR) image features with an assumption that LR image patches have the same sparse representations as their HR versions do. [19], [6] and [8] further improved Yang's

☆ This paper has been recommended for acceptance by Dr. A. Fernandez-Caballero

* Corresponding author: Tel.: +82 54 279 2881.
E-mail address: ath84@postech.ac.kr, atsj@nate.com (T.-H. An).

http://dx.doi.org/10.1016/j.patrec.2015.11.009
0167-8655/

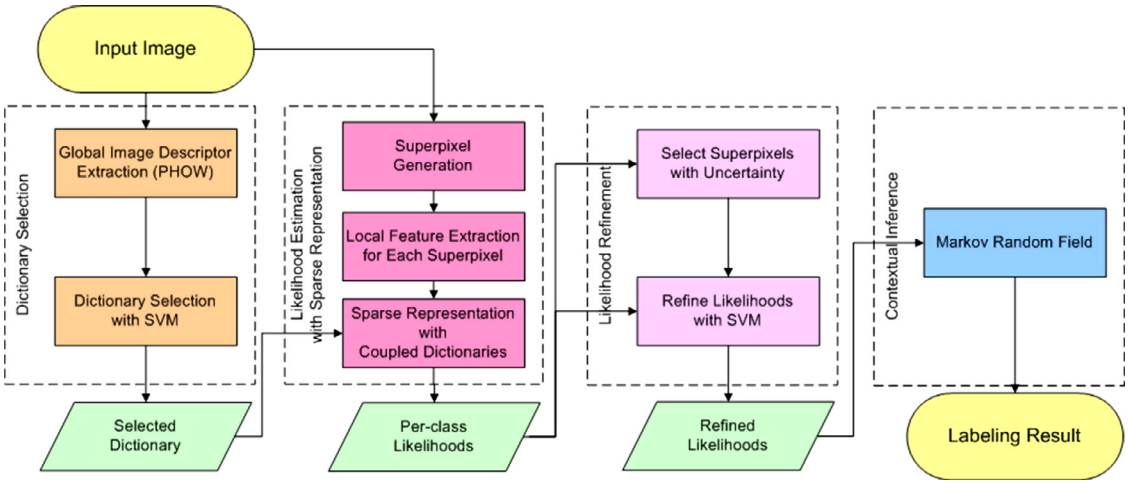

Fig. 1. Overall procedure.

joint dictionary learning scheme, and these methods have been successfully applied to super-resolution.

In this paper, we make the following contributions: (i) We propose a novel image parsing scheme based on sparse representation, and this is the first approach that utilizes sparse representation to the features of superpixels; (ii) We adopt a boosting concept approach to refine the likelihood for each superpixel; (iii) We reduce the data capacity comparing with the nonparametric scene parsing algorithms although the proposed algorithm has similar nature to that of nonparametric algorithms based on NN.

Fig. 1 shows the overall procedure of the proposed algorithm. The input is a still image. In order to select appropriate dictionary which is suitable to describe the current input image, the system extracts a global image descriptor first. Then a support vector machine (SVM) is used with the global image descriptor (Section 2.1). To extract local features, superpixels are generated using a graph-based segmentation algorithm (Section 2.2). Color, texture, location, shape of superpixels are used to obtain likelihoods by using sparse representation (Section 2.3). After obtaining likelihoods, superpixel-wise SVM score is used to refine them (Section 2.4). Markov random field (MRF) analysis is performed to incorporate neighborhood context (Section 2.5). These processes yield label information for each superpixel.

2. Approach

2.1. Dictionary selection

In nonparametric scene parsing approaches, image retrieval which finds images that are similar to the input image is an important and critical step since it determines the label candidates to parse the input image later. Because we are going to use sparse representation which requires trained dictionaries, this retrieval step is unsuitable for our case. Instead of image retrieval, we categorized training dataset images into several categories (e.g.: coast, street, forest) and the input image is classified into one of those categories. We use a variant of dense SIFT descriptors named PHOW [1] to describe the global image and train SVM with pre-defined image categories. For an image I_m, $S(m) \in 1, 2, \ldots, N_D$ and $D_{S(m)}$ indicate the selected image category and corresponding dictionary, with N_D as the number of dictionaries equal to the number of image categories, respectively.

2.2. Local feature extraction

Although we want to assign semantic labels to every pixel of the input image, a single pixel alone does not contain sufficient information for labeling. Thus we consider larger regions, i.e. superpixels and features from them. For an image, we extract superpixels with a fast graph-based segmentation algorithm [5]. Similar to [15], we represent each superpixel as a 794-dimensional vector which consists of appearances, color, texture, location and shape features. We compute color histograms by quantizing each R, G, B color space into 16 bins (48 dimensions). To represent texture information, a Histogram of maximum responses for 15 derivatives of oriented Gaussian filtered results (15 dimensions), a histogram of dense SIFT descriptor (dSIFT) for each superpixel (100 dimensions), a SIFT histogram of dSIFT for each dilated superpixel (100 dimensions) and four dSIFT histograms of left, right, top, and bottom boundary regions for each superpixel (4×100 dimensions) are used. We also compute a location histogram by quantizing the (x, y)-locations into a 8×8 grid (64 dimensions) and the top height of each superpixel relative to the image height (1 dimension). The shape of each superpixel is represented using its bounding box. We compute 8×8 mask of the superpixel over its bounding box (64 dimensions) and width/height of the bounding box relative to the image width and height (2 dimensions).

2.3. Likelihood estimation

The output of the scene parsing for an image I_m is a labeling $L = (l_1, l_2, \ldots, l_{N_s})$ which assigns a unique label $l_i \in 1, 2, \ldots, N_L$ to each superpixel s_i, where N_L and N_s are the total number of the semantic labels and superpixels respectively in an image.

2.3.1. Label transfer via sparse representation

Let $P_{LT}(a_i|l_i)$ be the appearance likelihood where LT denotes label transfer and a_i is the appearance feature vector described above. Conceptual illustration of likelihood estimation is depicted in Fig. 2. We compute $P_{LT}(a_i|l_i)$ with a weighted linear combination of a few bases, i.e. sparse representation.

$$\vec{P}_{LT}(a_i|l_i) = \begin{pmatrix} P_{LT}(a_i|l_i = 1) \\ P_{LT}(a_i|l_i = 2) \\ \vdots \\ P_{LT}(a_i|l_i = N_L) \end{pmatrix} = D^l_{S(m)} z_i \tag{1}$$

where $D^l_{S(m)}$ denotes the selected dictionary for the likelihood part and z_i is sparse coefficients estimated from

$$\min_{\{z_i\}_{i=1}^{N_s}} \sum_{i=1}^{N_s} \|a_i - D^a_{S(m)} z_i\|_2^2 + \gamma \|z_i\|_1 \tag{2}$$

and $D^a_{S(m)}$ is the selected dictionary for appearance part which is coupled with $D^l_{S(m)}$. Since $S(m)$ indicates the selected image category