A New Fast Approach to Nonparametric Scene Parsing

Parvin Razzaghi$^1$, Shadrokh Samavi$^{1,2}$

$^1$ Electrical and Computer Engineering Department, Isfahan University of Technology, Isfahan, Iran
$^2$ Electrical and Computer Engineering Department, McMaster University, Hamilton, Canada

Abstract

Scene parsing is a challenging research area in computer vision, which provides a semantic label for each pixel of the image. Most scene parsing approaches are parametric based which need models that are acquired through learning. 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 is proposed which transfers semantic labels from a training image to an unlabelled test image. 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 and Stanford background datasets. The obtained results show that our approach outperforms comparable state-of-the-art nonparametric approaches.

Keywords: Nonparametric scene parsing, label transferring, high level information, graphical model.

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 neighbouring 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 (Gonfaus et al., 2010, Kohli et al., 2007, Kohli et al., 2008, Ladicky et al., 2010), incorporating object detection results (Arbelaez et al., 2012, Choi et al., 2012, Ladicky, 2011, Yao et al., 2012) or incorporating context information (Cao and Fei-Fei, 2007, Lazebnik et al., 2006a, MICSUK and Kosecka, 2009, Su and Jurie, 2011).

In nonparametric approaches, instead of learning sophisticated models, knowledge from the labelled 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 approaches 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. (Liu et al., 2009) use SIFT flow (Liu et al., 2008) 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 establishes. Zhang et al. (Zhang et al., 2010) use KNN-MRF matching schema to find a correspondence map. To do this, for each test image, some small set of 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 & Zhang (Gould and Zhang, 2012) construct a graph of patch correspondences by using all images of the dataset during the training phase. To do this, PatchMatch algorithm (Barnes et al., 2009) 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 set) of dataset. If PatchMatchGraph were only constructed for the training images, then in the test phase, pairwise patch correspondence would have been needed. All of these approaches only transfer the label information. Recent nonparametric approaches try to employ contextual information to obtain more accurate results. Myeong et al. (Myeong et al., 2012) introduce a new data driven approach in which contextual relationships between all pair of regions in the labelled image are transferred to the unlabelled image by applying link analysis technique (Lu and Ip, 2010). Myeong and Lee (Myeong and Lee, 2013) add a new term to the CRF. Data and smoothness terms are computed based on conventional nonparametric approach (Tighe and Lazebnik, 2010, Tighe and Lazebnik, 2011). In a new term, high order semantic relations between objects are transferred from labelled 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. 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 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 labelling. 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 contribution of this paper is 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 and Stanford background datasets. Concluding remarks are given in section 5.

2 Transfer semantic labels

In this section, a generative approach to transfer semantic label from one labelled image to an unlabelled image is proposed. In the following, at first, formulation of the problem 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 $I$ as a training image in which each pixel is labelled. Each pixel is assigned one semantic label $l \in \{1,2,...,L\}$. Let $B$ be the number of class labels in image $I$. The regions of image $I$ are represented by $R = \{r_i\}_{i=1}^M$ and semantic label of each region is shown by $l(r_i)$. Let $M$ be the number of regions in image $I$. 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 $g: R \rightarrow R'$ assign each region in image $I$ to one region in image $I'$. Then, the label of the corresponding region in image $I$ is transferred to a region in image $I'$. We use a generative approach to determine function $g(.)$. It should be noted that, groups of regions in image $I$ can be assigned to one region in image $I'$. Conceptually, regions in test image $I$ are grouped to form one semantically meaningful unit. To obtain $g(.)$, a generative approach is proposed. Hence, we have:

$$g' = \arg \max_{s} p(g \mid I; I') = \arg \max_{s} \prod_{i} p(g_i \mid r_i; I') \cdot \prod_{i} \arg \max_{l} p(r_i \mid g; I') p(g; I')$$

where $g_i$ is the short form for $g(r_i)$ and it denotes corresponding region of $r_i$ in image $I$. It is noted that $g_i \in \{r_i\}_{i=1}^N$. In generative approaches, likelihood probability $p(r_i \mid g; I')$ and prior probability $p(g; I')$ are directly modelled. In our approach prior probability $p(g; I')$ is assumed to be uniform. Also, to model $p(r_i \mid g; I')$, a graphical model is constructed. In the following, how to model the likelihood probability is explained. Likelihood probability modelling is motivated from (Kim et al., 2010) in which a nonparametric approach to interactive segmentation is introduced.

2.1 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 labelled 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 (Arbelaez et al., 2011). 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. Therefore an undirected graph $G = (V, E)$ is constructed where nodes are shown by $V = \{V^t \cup V^q\}$ and the three types of edges are represented by $E = \{E^t, E^q, E^aq\}$. Edges are obtained by the following rules:

1) In type $E^t$ two regions in the training image $I^t$ are fully connected. These edges help to facilitate the transfer of spatial layout of semantic labels and their spatial extent to the test image. In another words, these edges encourage preserving the relative position of semantic labels in image $I$ with respect to each other.

2) In type $E^q$ two adjacent regions in the test image $I$ are connected. These edges are helpful in encouraging that neighbouring regions with a similar appearance to have the same semantic label.

3) In type $E^aq$ the connections between regions in images $I$ and $I'$ are considered. These edges provide to transfer appearance information and object level information from the training image to the test image.
2.3 Likelihood estimation

In this section, by using the relationships between nodes in the defined graphical model, a quadratic cost function is defined over regions likelihood. In the following, region likelihood \( p(r_i | r_j'; I') \) is represented by \( p_{ij} \). Let \( P \) denote a \( N \times M \) matrix representing regions likelihoods (\( p_{ij} \)) in image \( I \). It is noted that, \( P_{i,j}^{th} \) column of matrix \( P \), shown as \( P_i \), indicates degree of correspondence of regions of image \( I \) to the \( i^{th} \) region of image \( I' \). Finally, the cost function is defined as follows:

\[
J = e_s + e_s + e_c = \sum_i P_i - W_i \sum_j (P_i W_j - W_j) + \sum_j ((P_j C_i - C_j) (P_j C_i - C_j)') \n\]

where \( e_s, e_s \) and \( e_c \) are respectively defined as the smoothness, appearance and context terms. In the following, each of these terms is explained in details.

**Smoothness term (\( \sum_i P_i - W_i \)):** This term encourages similar and adjacent regions in image \( I \) to be grouped to form the same unit in image \( I' \). Let \( W \) denotes a \( N \times N \) matrix where each entry is computed as follows:

\[
w_{ij} = \begin{cases} \exp(-\alpha \Vert (r_i) - c(r_j) \Vert^2) & \text{if } r_i \in N_{ij} \\ 0 & \text{otherwise} \end{cases}
\]

where \( c(r_i) \) and \( c(r_j) \) respectively denote the mean color value of regions \( r_i \) and \( r_j \) in the Lab color space. Let \( N_{ij} \) indicates the neighbouring regions of region \( r_i \). Also, \( I_{ij} \) shows a \( N \times N \) identity matrix.

**Appearance term (\( \sum_i (P_i W_i - W_i) (P_i W_i - W_i)^T \)):** In this term similarity between appearances of groups of regions in image \( I \) with regions in the training image \( I' \) is considered. Let \( W^h \) denotes a \( N \times F \) matrix which \( i^{th} \) row indicates the descriptor of region \( r_i \) in the test image. Let \( F \) indicates the size of the descriptor of each region. Also, \( W^h \) denotes a \( M \times F \) matrix which \( j^{th} \) row of this matrix, as shown by \( W^h \), is the descriptor of region \( r_j \) in the training image.

Regions in the training image \( I' \) are semantically meaningful (i.e. they are objects). Hence, for a better similarity comparison, regions in image \( I \) should also be grouped to form semantically meaningful units. Similarity comparison is performed by comparing group of regions descriptors.
of image I with descriptors of corresponding region in image I'. Hence, $P_i^TW^h$ shows the descriptor of group of regions in image I which this group of regions corresponds with $i^\text{th}$ region of image I'. It combines descriptors of regions in a probabilistic framework by using $i^\text{th}$ column of the probability matrix $P$. In another words, $P_i^TW^h$ encodes object-level information. It should be noted that a descriptor of group of regions which is obtained by $P_i^TW^h$ should be the same as the descriptor that is achieved by directly describing group of regions. This property can be achieved by BoW model and hence we use BoW model to describe each region.

**Context term** ($\sum((P_i^TC_i - C^TP) - C_i^*)((P_i^TC_i - C^TP) - C_i^*)): This term encourages the transfer of spatial layout of semantic labels and their spatial extent in training image to the test image. Let $C^*$ denotes a $M\times M$ matrix which each entry of this matrix is defined as follows:

$$c_{ij}^* = O(r_i^*) - O(r_j^*)$$

where $O(r_i^*)$ and $O(r_j^*)$ indicate the centers of regions $r_i^*$ and $r_j^*$ in image I', respectively. In other words, matrix $C^*$ encodes the relative positions of regions in image I' with respect to each other. It should be noticed that the x component of the centroid is omitted since experimental results show that it does not carry any discriminative information in our approach. Let $C$ denotes a $N\times M$ matrix which is defined as follows:

$$C = [c_{n1} \ c_{n2} \ ... \ c_{nM}]$$

$$c_{ni} = [O(r_1) \ O(r_2) \ ... \ O(r_n)]^T$$

where $O(r_i)$ indicates the center of region $r_i$ in image I. The $i^\text{th}$ columns of the matrices $C^*$ and $C$ are indicated by $C^*_i$ and $C_i$.

Hence, the expression $(P_i^TC_i - C^TP)$ simulates the spatial layout of groups of regions with respect to the $i^\text{th}$ group of regions in the unlabelled image I. Therefore, the magnitude of the context term $c^*_C$ shows the similarity between the spatial layouts of semantic labels in the unlabelled image I and the labelled image I'.

Finally, likelihood probability of each region is estimated by minimizing the cost function in Equation 1. In the next subsection, optimization of the cost function is investigated.

### 2.4 Convex optimization

In this subsection, the cost function is converted to the standard form of scalar quadratic function and then optimized. Hence, at first, by modifying of matrices, it is converted as follows:

$$J = \tilde{P}^T(I_{N\times M} - \tilde{W})\tilde{P} + \tilde{P}^T\tilde{W}^\alpha + \tilde{P}^T\tilde{W}^\alpha^T - \tilde{W}^\alpha^T + \tilde{P}^T - \tilde{W}^\alpha^T$$

where the new notations of modified matrices are represented by $\tilde{}$. Definitions of these modified matrices are given in the Appendix. Also, $I_{N\times M}$ shows a $(N\times M)\times(N\times M)$ identity matrix. By rearranging Equation 2, it is converted to the standard form of scalar quadratic function:

$$J = \tilde{P}^T((I_{N\times M} - \tilde{W}) + \tilde{W}^\alpha(\tilde{W}^\alpha)^T + \tilde{C}_i\tilde{C}_i^T - \tilde{C}_i\tilde{C}_i^T + \tilde{C}_i\tilde{C}_i^T + \tilde{C}_i\tilde{C}_i^T)\tilde{P} +$$

$$(-2\tilde{W}^\alpha(\tilde{W}^\alpha)^T + 2\tilde{C}^\alpha\tilde{C}_i^\alpha - 2\tilde{C}^\alpha\tilde{C}_i^\alpha)^T + \tilde{C}^\alpha(\tilde{C}^\alpha)^T + \tilde{W}^\alpha(\tilde{W}^\alpha)^T$$

A standard scalar quadratic function $(x^TQx + bx + c)$ is convex if and only if $Q$ is positive semidefinite. Also, for any real matrix $Z$, it is guaranteed that $ZZ^T$ is positive semidefinite (Gärtnert and Matousek, 2012). Hence, to guarantee that Equation 3 is convex, it is rewritten as follows:

$$J = \tilde{P}^T(ZZ^T)\tilde{P} + (-2\tilde{W}^\alpha(\tilde{W}^\alpha)^T + 2\tilde{C}^\alpha\tilde{C}_i^\alpha - 2\tilde{C}^\alpha\tilde{C}_i^\alpha)^T + \tilde{C}^\alpha(\tilde{C}^\alpha)^T + \tilde{W}^\alpha(\tilde{W}^\alpha)^T$$

where $Z$ is defined as:

$$Z = (I_{N\times M} - \tilde{W}) + \tilde{W}^\alpha(\tilde{W}^\alpha)^T + \tilde{C}^\alpha\tilde{C}_i^\alpha - \tilde{C}_i\tilde{C}_i^\alpha + \tilde{C}_i\tilde{C}_i^\alpha$$

$$Z = (I_{N\times M} - \tilde{W}) + \tilde{W}^\alpha(\tilde{W}^\alpha)^T + \tilde{C}^\alpha\tilde{C}_i^\alpha - \tilde{C}_i\tilde{C}_i^\alpha + \tilde{C}_i\tilde{C}_i^\alpha$$
Finally, the formulation of the problem would be as follows:

$$\min_J \quad J$$

$$\text{s.t.} \quad 0 \leq p_i \leq 1$$

$$\forall i = 1, ..., N, \ j = 1, ..., M$$

The modified cost function with constraint is convex. Hence, we use a standard optimization package called CVX (Grant and Boyd, 2012) to solve this problem. Now, for each region in image $I$, $M$ probabilities (i.e. the rows of matrix $P$) are associated which denotes the degree of belonging it to each region in image $I$. Finally, the semantic label of each region in image $I$ is achieved by maximum a posterior probability (MAP) estimation. However, as it is mentioned, prior probability is assumed to be uniform. Therefore, MAP estimation is reduced to maximum likelihood (ML) estimation. Hence, we have:

$$g' = \arg \max_{g \in \{g^l\}_{l=1}^N} p(r_i \mid g); I)$$

(7)

where the corresponding region of $r_i$ in image $I$ is $g_i$. Therefore, each region in image $I$ is assigned to a region in image $I$ by function $g'$. Finally, for labelling regions of image $I$, semantic labels of corresponding regions in image $I$ are transferred to regions in image $I$. Hence the semantic label of region $r_i$ in image $I$ is $l(g_i')$. In our approach an additional label $l_F$ is given which considers situation that a region in the test image should not be assigned to any region in image $I$. Region $r_i$ in image $I$ is assigned to $l_F$ when $p(r_i \mid g_i'); I)$ is below the 0.5. Semantic label is not assigned to regions which is corresponded to $l_F$.

It should be noted that, our proposed method to transfer knowledge from training sample to the test sample does not require patch correspondence and is solved by optimizing a quadratic convex function. Hence, it is fast.

### 3 Scene Parsing

In this section, a new nonparametric and fast scene parsing approach is proposed. To do so, at first, for each test image, some similar training images are retrieved. To retrieve images, spatial pyramid matching (SPM) (Lazebnik et al., 2006b) is used. In our approach SIFT (Lowe, 2004), Hue descriptor (Weijer and Schmid, 2006) and 17-filter bank are used as descriptors. Let $\{I^t\}_{t=1}^T$ denotes the top $k$ retrieved images for test image $I$. Then, for a set of retrieved images, we pair the test image with each of the retrieved training images. Then for each pair, based on the proposed generative label transfer approach, the knowledge of the training image is transferred to the test image. Hence, we have:

$$g' = \arg \max_{g \in \{g^l\}_{l=1}^N} p(g \mid I; I') = \arg \max_{g \in \{g^l\}_{l=1}^N} \prod_{i=1}^{m} p(g_i \mid r_i; I') = \prod_{i=1}^{m} \alpha \max_{g_i \in \{g^l\}_{l=1}^N} p(r_i \mid g_i'); I)p(g_i; I'), \ \forall t = 1, k$$

where $g' = \{g_i^l\}_{l=1}^N$ represents the corresponding regions of image $I$ in image $I'$. Finally, to labelling each pixel of the test image, the transferred information of the training images are aggregated using conditional random field (CRF). Hence, the energy function for pixel labelling of image $I'$ is defined as follows:

$$E(L) = \sum_{l \in L} \psi_d(l) + \sum_{l, j \in L} \psi_g(l, l_j)$$

(8)

where $\psi_d(l)$ is a data term which represents the cost of assigning label $l$ to region $r_i$ and $\psi_g(l_i, l_j)$ is a smoothness term which encourages similar neighbouring regions to have the same label. In the following, these terms are defined. **Data term:** In the data term, similarity appearance between corresponding regions of the training images and the test image is considered. To do this, three appearance cues are used: BoW model, texture and color. Hence, data term in CRF model is defined as follows:

$$\psi_d(l_i) = \begin{cases} \min_{g \in \{g^l\}_{l=1}^N} \max_i \left[ d_i(r_i, g_i'), d_i(r_i, g_i'), d_i(r_i, g_i') \right] & \text{if } l_i \in \{l(g^l_i)\}_{l=1}^N \\ 0 & \text{otherwise} \end{cases}$$

(9)

where $d_i(r_i, g_i')$ shows the dissimilarity between BoW histograms of corresponding regions.
Therefore, it is defined as follows:

\[ d_i(r_i, g'_i) = 1 - \exp(-0.5 \times \|BoW(r_i) - BoW(g'_i)\|^2) \]

where \( BoW(r_i) \) denotes the BoW histogram of region \( r_i \). Here, to describe region’s texture, Local Binary Pattern (LBP) is used (Ojala et al., 2002). Hence, for each pixel uniform LBP is computed and then histogram of uniform LBP is calculated for pixels inside region. Hence, to obtain texture dissimilarity between corresponding regions, we have:

\[ d_z(r_i, g'_i) = 1 - \exp(-0.5 \times \|LBP(r_i) - LBP(g'_i)\|^2) \]

where \( LBP(r_i) \) denotes the histogram of region \( r_i \). Finally, for color descriptor, we have:

\[ d_z(r_i, g'_i) = 1 - \exp(-0.5 \times \|c(r_i) - c(g'_i)\|^2) \]

where \( c(r_i) \) and \( c(g'_i) \) denotes the mean color value of region \( r_i \) and \( g'_i \) respectively. In Equation 9, three appearance cues are combined using max operator. This causes error in one of appearance cues (BoW model, texture and color) is compensated by another one which reduces the false positive case of the method. Eventually, \( \psi(l) \) is obtained using min operator over dissimilarity of different corresponding regions which are obtained from different retrieved images.

**Smoothness term:** it encourages the neighboring regions to have the same labels. In this paper, the smoothness term is defined based on the color differences between neighboring regions:

\[ \psi_s(l_i, l_j) = \begin{cases} 
\exp\left(-\frac{\|c(r_i) - c(r_j)\|^2}{2\beta}\right) & \text{if } r_i \in N_i \land l_i \neq l_j \\
0 & \text{otherwise}
\end{cases} \]

where \( \beta \) is the mean square difference between mean color value of all adjacent regions of the image. Finally, the energy function is optimized by Graph cut algorithm (Boykov et al., 2001, Kolmogorov and R.Zabih, 2004, Boykov and Kolmogorov, 2004). Pseudo code of the proposed algorithm is given in Algorithm 1.
In order to evaluate the proposed approach, it is applied on two well-known datasets of MSRC-21 (Shotton et al., 2008) and Stanford background (Gould et al., 2009). In the proposed approach, the number of retrieved images $k$ is set to 5. To retrieve images, SPM is used. To do this, key points are densely sampled which sampling step is considered 8 pixels in horizontal and vertical directions. To describe each key point SIFT (Lowe, 2004), Hue descriptor (Weijer and Schmid, 2006) and 17-filter bank are used. For each descriptor, a 200 visual word dictionary is constructed independently. Then, the BoW histogram of each descriptor is concatenated to form the final descriptor of each image. It should be noted that the obtained visual words in the retrieval step are used to describe each region, too.

**MSRC-21 dataset**: it contains 21 categories of classes. Images are divided into train and test sets based on standard split method (Shotton et al., 2008). To show the performance of the proposed approach, the number of retrieved images $k$ is set to 5. To retrieve images, SPM is used. To do this, key points are densely sampled which sampling step is considered 8 pixels in horizontal and vertical directions. To describe each key point SIFT (Lowe, 2004), Hue descriptor (Weijer and Schmid, 2006) and 17-filter bank are used. For each descriptor, a 200 visual word dictionary is constructed independently. Then, the BoW histogram of each descriptor is concatenated to form the final descriptor of each image. It should be noted that the obtained visual words in the retrieval step are used to describe each region, too.

**Algorithm 1**

**Knowledge _Transfer (I', I'_0, I)**

**Input:**
- $I'$: Training image
- $I'_0$: Ground truth of the training image
- $I$: Test image

**Output:**
- $g': R \rightarrow R': Assign each region in image $I$ to one region in image $I'$

**Nonparametric Scene Parsing**

**Input:**
- $\{I'_t\}_{t=1}^M$: set of training images
- $\{I'_t\}_{t=1}^M$: set of ground truth of the training image
- $M'$: number of training images
- $I$: test image

**Output:**
- $L = \{l_i\}_{i=1}^N$: Semantic label of regions of image $I$

**Algorithm 1 - Pseudo code of the proposed algorithm.**

**4 Experimental Results**
method of transferring knowledge from one training image to test image by optimizing Equation 6, some examples are shown in Figure 2. Black regions denote that these regions are not assigned to any region of training image and hence are assigned to $l_p$. Therefore, semantic labels are not assigned to these regions. Such situation occurs when the desirable label of test image is not observed in training image (Figure 2-b) or the probability of assigning region to semantic labels is below the 0.5 (Figure 2-a). Also, in some cases, labels are wrongly transferred due to appearance similarities (Figure 2-c). It is noted that in scene parsing, CRF is used to aggregate the obtained information from $k$ retrieved images, hence, some of these wrong assignments are omitted.

The achieved results on MSRC-21 dataset are shown in Table 1. Our proposed method is compared with the state-of-the-art parametric and nonparametric approaches in Table 1. As it is shown, our method achieves 2.1% gain in accuracy on the state-of-the-art nonparametric approaches on MSRC dataset. In (Gould and Zhang, 2012), in the training stage by using all data (training and test data), a PatchMatchGraph is learned. It takes 1 hour and 32 minutes for $k=5$ and 4 hours for $k=10$ with 24 GB of memory system. It should be noticed that our proposed approach does not require any learning in the training stage. Also, our method, unlike the other nonparametric approaches, does not require patch correspondence, neither in the train nor in the test stages. This has given a low runtime property to our method. A plot of inference time of the proposed generative approach for label transferring of 1200 different image pairs is shown in Figure 3. It shows that our proposed approach takes an average of 0.2 seconds for transferring knowledge from a training image to a test image. It confirms that our proposed method is very fast. It should be noticed that our MATLAB code is run on a single core of 3.2 GHz Intel CPU with 4 GB RAM.

In comparison with parametric approaches, our proposed method outperforms (Shotton et al., 2006) and has comparable performance with (Gonfaus et al., 2010) and it is below (Yao et al., 2012). However, it should be noted that our method is nonparametric. Parametric approaches are commonly complicated and use heavily trained discriminative models. Also, for any specific dataset, parametric approaches need to learn a new model. However, nonparametric approaches are simple and their generalization capability is high. Hence, they are applicable to any new dataset without the need for learning of a new model. Some qualitative results on MSRC-21 dataset are shown in Figure 4.

The retrieval step has a significant impact on final results. To do this, an experiment is designed to investigate whether all semantic labels in the test image have been observed in the set of retrieved images or not. The achieved results show that in 83% of cases, all semantic labels in test images are observed in the set of retrieved images. It is clear that if the retrieval step is done better, the performance of the proposed approach is increased.

### Table 1: MSR-21 segmentation results

<table>
<thead>
<tr>
<th>Approach</th>
<th>Our, $k=5$</th>
<th>(Gould and Zhang, 2012), $k=10$</th>
<th>(Gould and Zhang, 2012), $k=20$</th>
<th>(Jain et al., 2010)</th>
<th>(Shotton et al., 2006)</th>
<th>(Gonfaus et al., 2010)</th>
<th>(Yao et al., 2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>74.6</td>
<td>71.1</td>
<td>72.5</td>
<td>69.3</td>
<td>72.2</td>
<td>77</td>
<td>87</td>
</tr>
<tr>
<td>Mean</td>
<td>64.1</td>
<td>64.7</td>
<td>66.4</td>
<td>-</td>
<td>57.7</td>
<td>75</td>
<td>77</td>
</tr>
</tbody>
</table>
Figure 2- Some examples of transferring knowledge from one training image to the test image. (a) Labels are correctly transferred (b) the desirable label of test image is not observed in training image (c) labels are wrongly transferred.

Figure 3- Plot of inference time for 1200 different pair of images. Green dots denote the inference time of each pair and the red line corresponds to the mean of inference times.
Stanford background dataset: it contains 715 images which contains 8 semantic class labels (Gould et al., 2009). It is shown that in the retrieval step, in 90% of the images in the test set, all semantic labels in the test image are shown in the set of retrieved images. The obtained results on this dataset are shown in Table 2. In all categories, our approach receives the acceptable accuracy except for the “mountain” class. Because, in 40% cases of test images, ‘mountain’ class label is not observed in the set of retrieved images. Also, it is noted that in case of small objects, there are few key points inside small objects. In some cases where the object size is lower than the sampling step, it is possible that no key point exists inside it. As a result, these regions cannot be assigned to a true region in the transferring knowledge step. Hence, it causes the performance of the proposed approach to decrease. Our approach is compared with the state-of-the-art methods in Table 2. As it is shown, our approach outperforms the state-of-the-art nonparametric methods. Tighe and Lazebnik (Tighe and Lazebnik, 2010), to obtain an initial labelling for each pixel, use trained boosted decision tree classifier instead of superpixel correspondence. This is in contrast to our approach which does not require any learning based model. Also, in comparison with parametric approaches, our approach outperforms (Gould et al., 2009) and has a comparable performance with (Gould, 2012). Table 2 shows the ability of the proposed approach which with simplicity achieves the strong performance. While our approach, in comparison with the other methods, is fast. Some qualitative results on Stanford dataset are shown in Figure 5.

Table 2- Stanford background segmentation results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Building</th>
<th>Grass</th>
<th>Tree</th>
<th>Sky</th>
<th>Water</th>
<th>Road</th>
<th>Mountain</th>
<th>Foreground</th>
<th>Global</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our</td>
<td>75.2</td>
<td>91.4</td>
<td>71.9</td>
<td>94.7</td>
<td>80.8</td>
<td>86.3</td>
<td>13.1</td>
<td>50.7</td>
<td>78.46</td>
<td>70.51</td>
</tr>
</tbody>
</table>

Comparison of our results with the published state-of-the-art results

<table>
<thead>
<tr>
<th>Type</th>
<th>Nonparametric</th>
<th>Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>78.46</td>
<td>70.7</td>
</tr>
<tr>
<td>Mean</td>
<td>70.51</td>
<td>57.0</td>
</tr>
</tbody>
</table>
In this paper, we propose a new fast nonparametric approach for scene parsing. All nonparametric scene parsing approaches require pairwise patch correspondences. Our proposed approach does not require patch correspondences, neither for the training and nor for the test phase. Hence, our approach is fast and is not limited to any specific dataset. The proposed approach outperforms many of the state-of-the-art nonparametric methods. Also, our nonparametric approach is comparable to some successful and recent parametric approaches. Our method in comparison to some other nonparametric approaches does not use any learning based model to improve its accuracy. In scene parsing approaches, involving of high order information is an important step. We use both object-level information and contextual information.

To evaluate the proposed approach in considering object level information, it is applied to the MSRC-21 dataset. Results showed that the proposed approach achieved 81.2% accuracy on non-object classes and 60.6% accuracy on object classes (excluding the boat class). Only a few nonparametric methods have demonstrated the obtained accuracy on this dataset. Since some of the approaches which have used this dataset, have not given their per class accuracies, we could not compare our per class accuracies. In the case of small objects, these objects could not be segmented into distinct segments. Hence, these objects were not assigned true semantic labels. Also, in some cases which small objects were segmented into distinct segments, there was the possibility that no key point existed inside them. Therefore, in the process of transferring knowledge from one training image to a test image, those segments could not be assigned to a true segment in the training image.

## Appendix

The modified matrices of the Equation 2 are defined as follows:

$$\hat{p} = \begin{bmatrix} p_{11} & p_{21} & \ldots & p_{n1} & p_{12} & \ldots & p_{n2} & \ldots & p_{1m} & \ldots & p_{nm} \end{bmatrix}^T$$

$$\hat{W} = \begin{bmatrix} W_{n \times n} & 0_{n \times n} & \ldots & 0_{n \times n} \\ 0_{n \times n} & W_{n \times n} & \ldots & 0_{n \times n} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n \times n} & \ldots & \ldots & W_{n \times n} \end{bmatrix}_{(N \times M)^2}$$

Figure 5- some qualitative results of the proposed approach on Stanford background dataset. (a) The input image (b) segmentation results (c) ground truth.

## 5 Conclusions

In this paper, we proposed a new fast nonparametric approach for scene parsing. All nonparametric scene parsing approaches require pairwise patch correspondences. Our proposed approach does not require patch correspondences, neither for the training and nor for the test phases. Hence, our approach is fast and is not limited to any specific dataset. The proposed approach outperforms many of the state-of-the-art nonparametric methods. Also, our nonparametric approach is comparable to some successful and recent parametric approaches. Our method in comparison to some other nonparametric approaches does not use any learning based model to improve its accuracy. In scene parsing approaches, involving of high order information is an important step. We use both object-level information and contextual information. To evaluate the proposed approach in considering object level information, it is applied to the MSRC-21 dataset. Results showed that the proposed approach achieved 81.2% accuracy on non-object classes and 60.6% accuracy on object classes (excluding the boat class). Only a few nonparametric methods have demonstrated the obtained accuracy on this dataset. Since some of the approaches which have used this dataset, have not given their per class accuracies, we could not compare our per class accuracies. In the case of small objects, these objects could not be segmented into distinct segments. Hence, these objects were not assigned true semantic labels. Also, in some cases which small objects were segmented into distinct segments, there was the possibility that no key point existed inside them. Therefore, in the process of transferring knowledge from one training image to a test image, those segments could not be assigned to a true segment in the training image.
\[
\tilde{W}^h = \begin{bmatrix}
W_{N\times F}^h & 0_{N\times F} & \cdots & 0_{N\times F} \\
0_{N\times F} & W_{N\times F}^h & \cdots & 0_{N\times F} \\
\cdots & \cdots & \cdots & \cdots \\
0_{N\times F} & \cdots & \cdots & W_{N\times F}^h
\end{bmatrix}_{(N\times M)\times (F\times M)}
\]

\[
\tilde{W}^* = [w_{11}^* \ldots w_{1p}^* \ldots w_{M1}^* \ldots w_{Mq}^*]_{(p\times M)}
\]

\[
\tilde{C}_1 = \begin{bmatrix}
C_{N\times M} & 0_{N\times M} & \cdots & 0_{N\times M} \\
0_{N\times M} & C_{N\times M} & \cdots & 0_{N\times M} \\
\cdots & \cdots & \cdots & \cdots \\
0_{N\times M} & \cdots & \cdots & C_{N\times M}
\end{bmatrix}_{(N\times M)\times (M\times M)}
\]

\[
\tilde{C}_2 = \begin{bmatrix}
C_{N\times d} & 0_{N\times d} & \cdots & 0_{N\times d} & C_{N\times d} & 0_{N\times d} & \cdots & 0_{N\times d} \\
0_{N\times d} & C_{N\times d} & \cdots & 0_{N\times d} & 0_{N\times d} & C_{N\times d} & \cdots & 0_{N\times d} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0_{N\times d} & \cdots & C_{N\times d} & 0_{N\times d} & \cdots & 0_{N\times d} & \cdots & C_{N\times d}
\end{bmatrix}_{(N\times M)\times (M\times M)}
\]

\[
\tilde{C}^* = [c_{11}^* \ldots c_{1q}^* \ldots c_{M1}^* \ldots c_{Mq}^*]_{(q\times M)}
\]

**References**


