Segment an Image by Looking into an Image Corpus

Xiaobai Liu[†,‡], Jiashi Feng[‡], Shuicheng Yan[‡], Liang Lin[§], Hai Jin[†]
[†] Huazhong University of Science and Technology, China,
[‡] National University of Singapore, Singapore
[§] Sun Yat-Sen University, Guangzhou, China

Abstract

This paper investigates how to segment an image into semantic regions by harnessing an unlabeled image corpus. First, the image segmentation task is recast as a small-size patch grouping problem. Then, we discover two novel patch-pair priors, namely the first-order patch-pair density prior *and the second-order* patch-pair co-occurrence prior*, founded on two statistical observations from the natural image corpus. The underlying rationalities are: 1) a patch-pair falling within the same object region generally has higher density than a patch-pair falling on different objects, and 2) two patch-pairs with high co-occurrence frequency are likely to bear similar semantic consistence confidences (SCCs), i.e. the confidence of the consisted two patches belonging to the same semantic concept. These two discriminative priors are further integrated into a unified objective function in order to augment the intrinsic patch-pair similarities, originally calculated using patch-level visual features, into the semantic consistence confidences. Nonnegative constraint is also imposed over the output variables and an efficient iterative procedure is provided to seek the optimal solution. The ultimate patch grouping is conducted by first building a similarity graph, which takes the atomic patches as vertices and the augmented patch-pair SCCs as edge weights, and then employing the popular Normalized Cut approach to group patches into semantic clusters. Extensive image segmentation experiments on two public databases clearly demonstrate the superiority of the proposed approach over various state-of-the-arts unsupervised image segmentation algorithms.*

1. Introduction

The task of segmenting semantic regions or parsing image is critical since a wide range of image related problems could in principle take full advantages of the semantically segmented images, such as content-based image retrieval, multi-label image annotation and part-based object recognition [4],[20, 6],[17]. Generally, for one given image, parsing it into semantically meaningful elements is not easy, due to the large intra-object variations. Extra information beyond individual image is intuitively valuable for solving this problem. With the proliferation of Internet images from popular photo sharing websites (e.g. Flickr) and image search engines (e.g., GOOGLE and BING), one promising direction is to explore the large-scale image corpus and discover valuable contextual information for boosting semantic image segmentation [7, 19]. One common strategy is to search for structurally similar images using certain effective feature descriptors and utilize the matches to facilitate image segmentation [15, 16]. Despite the impressive results and successes demonstrated in the past literature, for many types of images the segmentation quality is still not satisfying due to the facts that i) the low-level visual features may not be powerful enough to capture semantic similarity, and ii) it is not reasonable, or even impossible, to expect to always find structurally similar matches from an image corpus.

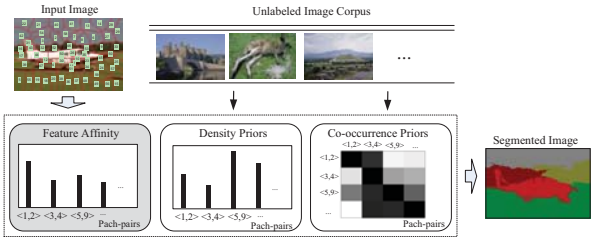

Figure 1. Schematic illustration of the proposed approach.

To tackle above issues, we present a *practical* approach to segment one given image by extracting implicit priors from an unlabeled image corpus. We first over-segment the input image into an ensemble of localized atomic patches and then cast the common image segmentation task as a patch grouping problem. Within this context, the intrinsic patch-pair affinity, which is originally calculated as the appearance similarity based on patch-level features, is augmented into the semantic consistence confidence (SCC), i.e., the confidence of these two patches belonging to the same semantic concept, by harnessing two statistical obser-

vations on the unlabeled image corpus. One is the *patch-pair density prior* [13]. We find out that a patch-pair falling within the same semantic object/class usually have higher density than a patch-pair consisting of two semantically different patches. The patch-pair densities can be directly calculated from the unlabeled image corpus by matching each patch-pair in the given image to visually similar patch-pairs in the unlabeled image corpus. This prior provides a discriminative measurement of semantic consistence for determining whether two patches belong to the same semantic object, and the results of our previous work [13] have shown its great potential with image segmentation task.

The other observation is the *patch-pair co-occurrence prior*. Similar to the density prior, it is also founded on a statistic from the unlabeled image corpus: if two patch-pairs frequently co-occur within short spatial distance, they are both likely to bear high semantic consistence confidences, namely, the two patch-pairs are semantically kindred with high probability at the same time. This observation is intuitively reasonable since the couples of intra-class patch-pairs generally have dominant population within the natural image corpus. Beyond the first-order density prior which describes the relationship of individual patch-pair itself, patch-pair co-occurrence prior further provide high-order information for describing the interdependence among patch-pairs, which can improve the distinctness and robustness of the desired patch-pair semantic consistence confidence.

Figure 1 demonstrates the entire flowchart of our proposed approach. For a given image, our approach combines three different sources of information for semantically patch grouping, including i) the intrinsic patch affinities, ii) the first-order patch-pair density prior, and iii) the second-order patch-pair co-occurrence prior. The first one is derived from the given image itself and serves as a low-level visual representation, while the latter two are derived from the auxiliary unlabeled image corpus and serve as high-level semantic representations. They are in essence complementary to each other, and thus the combination of them is expected to enhance the discriminating capability while describing the patch-pairs. We formulate the above three diverse cues into one unified objective function in order to seek the optimal semantic consistence confidence for every patch-pair. The patch-pairs with high confidences are most likely to contain semantically identical patches, and vice versa. We further utilize the patch-pair confidences to construct the augmented patch similarity graph, on which the popular spectral clustering technique, Normalized Cut [17], is employed to conduct semantic patch grouping.

It is worthwhile to highlight three aspects of the proposed approach here: i) the proposed patch-pair priors are simple yet effective in measuring the semantic consistences of the patch-pairs; 2) the proposed approach can boost single image segmentation accuracy with the help of auxiliary unlabeled image corpus under various conditions, e.g., it can work well even with a small-size image corpus and 3) no parametric models from the image corpus are learnt and thus it is scalable to harness a large-scale image corpus. In contrast with our previous work [13] that first presents the first-order patch-pair density prior, this work further proposes a second-order prior and formulate it within a unified formulation in the framework of nonnegativity analysis. These advantages and extensions shall be further validated by extensive experiments on publicly available image databases.

2. Relation with Previous Works

Our approach is closely related to the recent advances in the community of computer vision. In the literature, a broad family of approaches to image segmentation has been proposed. The typical ones include integrating features such as brightness, color, or texture over local image patches and then clustering those features based on fitting mixture model [21, 3], model-finding [4] or graph partitioning [20, 6]. Among them, three algorithms are most widely used in recent applications, including Tu and Zhu's data driven MCMC algorithm (DDMCMC) [20], Comaniciu and Meer's Mean Shift [4], and Shi and Malik's Normalized Cuts [17]. DDMCMC could achieve high performance but is usually computationally expensive. Mean Shift and Normalized Cuts are easy to implement and can provide reasonably good precision, but often produce artifacts by breaking large uniform regions into chunks. Moreover, these approaches only utilize the information contained in individual image itself, where the parsing task is painfully underconstrained and limited to the low segmentation accuracy.

Recently, with the increasing availability of Internet/Web image set, large database-driven approaches have shown the great potentials for nonparametric semantic image segmentations task. Shakhnarovich *et al*. [16] proposed to estimate the pose of human relying on 0.5 million training examples. Hays *et al*. [7] proposed to fill holes on an input image by introducing elements that are likely to be semantically correct through searching in a large image set. Torralba *et al*. [19] proposed to compute a simple sum of squared difference (SSD) match on the localized image parts, e.g., rectangles, yet obtained a semantically meaningful parsing. Similarly, Russell [15] proposed to partition the input image into composite regions and seek their matches from a large unlabeled image corpus, in order to explain the input image. Nevertheless, the major block of applying above algorithms is how to achieve a good tradeoff between the matching accuracy and computational cost. In contrast, our approach also uses the auxiliary image corpus to improve image segmentation accuracy but can work well on either small or middle size image set, benefiting from the discovered patch-pair priors

and the unified objective function which combines multiple complementary cues.

3. Segment Image with Patch-pair Priors from an Unlabeled Image Corpus

3.1. Cast Image Segmentation as Patch Grouping

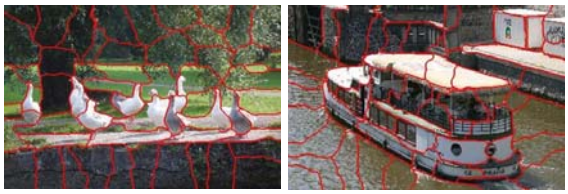

Figure 2. Exemplar images overlaid with over-segmentation results. Images shown are from the MSRC database [18].

We partition one image into localized atomic patches for image representation and thus the task of image segmentation is recast as grouping the atomic patches into larger-size semantic regions. Each atomic patch comprises pixels that are spatially coherent and perceptually similar with respect to certain appearance features, such as intensity, color and texture. Many algorithms have been proposed for image over-segmentation in the past literature, and in this work, we choose to use the method provided by Ren *et al*. in [14]. Figure 2 shows some image over-segmentation examples. We can observe that the localized patches are roughly homogeneous in size and shape, which simplifies the computation in later stages. Each image is resized with a ratio of $400/max(width, height)$ and partitioned into $40 \sim 50$ patches on average. On an Intel Xeon X5450 workstation with 3.0GHz CPU and 16GB memory, it takes about 60 seconds to process one image. A basic assumption here is that every atomic patch is roughly involved within one single object/class. Formally, let x_i denote the i-th patch within the given image or any unlabeled image, and $\mathbf{x}_i = \{x_{i_1}, x_{i_2}\}$ denote the corresponding pair of patches x_{i_1} and x_{i_2}. We describe each patch by a Local Binary Pattern (LBP) [1] descriptor, denoted as f_{x_i}. We group every two different patches within the same image to form one patch-pair from which one LBP feature is extracted and denoted as $f_{\mathbf{x}_i}$.

3.2. I: Intrinsic Patch-pair Affinity

The crucial step of patch grouping is to compute the semantic consistence confidence (SCC) for every patch-pair, i.e., the probability of belonging to the same semantic object/class, for every two patches. For a patch-pair $\mathbf{x}_i = \{x_{i_1}, x_{i_2}\}$, the semantic consistence confidence is closely related with patch-pair affinity w_i, which can be estimated based on the low-level visual features as follows,

$$w_i = w_{\{x_a, x_b\}} = \exp\{-\mathcal{D}(f_{x_a}, f_{x_b})/2\sigma^2\}, \quad (1)$$

where $\mathcal{D}(\cdot,\cdot)$ measures the Euclidean distance between two feature vectors and σ is the scale parameter. We fix σ to be 1 in this work. Patch-pairs with high affinities are likely to comprise semantically kindred patches. Let h_i denote the desired semantic consistence confidence for the patch-pair $\mathbf{x}_i$, we have following objective function to optimize:

$$\min_h \sum_i (h_i - w_i)^2 \quad s.t. \quad h \geq 0, \quad (2)$$

which imposes nonnegative constraint on the output variables in order to make the desired confidences more informative [10]. Herein, $h \in R^n$ is an n-dimensional vector and n indicates the number of patch-pairs within the given image.

3.3. II: First-order Patch-pair Density Prior

The patch-pair density prior is founded on the following observation: for natural images, semantically kindred patch-pairs generally possess higher densities than semantically inhomogeneous patch-pairs. The patch-pair density can be directly estimated from the unlabeled image corpus using certain density estimation method. Herein, we adopt the popular Parzen window [5] method. This observation shows that one patch-pair with low density tends to contain two semantically different patches, and vice verse. Formally, let $\mathbf{x}_i$ be one patch-pair from the input image and $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K\}$ denote its K-nearest patch-pairs retrieved from the unlabeled image corpus. We can estimate the density of $\mathbf{x}_i$, denoted as c_i, as follows,

$$c_i \propto \frac{1}{K} \sum_{l=1}^{K} e^{-\mathcal{D}^2(f_{\mathbf{x}_i}, f_{\mathbf{y}_l})}, \quad (3)$$

where $\mathcal{D}$ denotes the Euclidean distance between two feature vectors.

As indicated by the density prior, it is natural to enforce that patch-pairs with higher densities should also bear higher SCCs and vice versa. Formally, we can derive h_i for the patch-pair $\mathbf{x}_i$ from c_i as follows:

$$\min_h \sum_i h_i(1 - c_i) \quad s.t. \quad h \geq 0. \quad (4)$$

We verify the above prior on the image set from the MSRC [18] database, where each image is provided with segmentation groundtruth. We first partition the given image into atomic patches and associate each patch-pair with a label of 'inter' or 'intra'-object/class according to the groundtruth. Then, for each patch-pair, a 59-dimensional LBP descriptor is extracted and $K = 1000$ nearest patch-pairs are retrieved from the unlabeled image corpus based on Euclidean distance. Thus, we estimate the density of

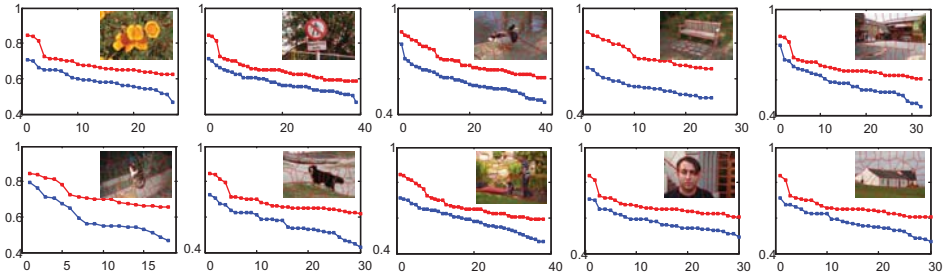

Figure 3. Density comparison between intra-class (red color) and inter-class (blue color) patch-pairs. The images are from the MSRC database [18]. See texts for detailed descriptions.

a patch-pair using the Parzen window method and normalize it by the maximum density value within the same image. We plot the normalized densities in descending order. Figure 3 shows the density comparison between intra-class patch-pairs (red color) and inter-class (blue color) patch-pairs in 10 images containing different semantic concepts. The horizontal axis represents the indices for the top patch-pair and the vertical axis represents their estimated densities. We can see that intra-class patch-pairs usually have much higher densities as compared to the inter-class ones, which demonstrates the discriminating power of the proposed patch-pair density prior.

3.4. III: Second-order Patch-pair Co-occurrence Prior

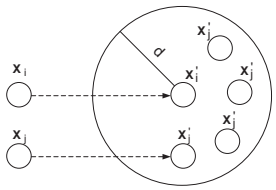

Figure 4. Spatial histogram.

The second-order patch-pair prior indicates that, if two patch-pairs from the given image co-occur with high frequency in an unlabeled image corpus, they are likely to bear similar semantic consistence confidences. For every two patch-pairs in the given image, we first compute their visually K-nearest patch-pairs from the unlabeled image corpus respectively, and then identify their co-occurring neighbors within a given spatial distance. We use the spatial histogram [11] for spatial relationships computation. As illustrated in Figure 4, let $\mathbf{x}'_i$ and $\mathbf{x}'_j$ denote the retrieved neighbors of the patch-pairs $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. The spatial histogram centered at $\mathbf{x}'_i$ with respect to patch-pair $\mathbf{x}_j$ is defined as $\mathcal{H}(\mathbf{x}'_i, \mathbf{x}_j)(d) = m_{k,j}$, where the spatial radius d defines the supporting region of the histogram and $m_{k,j}$ denotes the number of $\mathbf{x}'_j$ falling in the supporting region. The ultimate co-occurrence frequency of $\mathbf{x}_i$ and $\mathbf{x}_j$ is calculated as $\sum_{\mathbf{x}'_i} \mathcal{H}(\mathbf{x}'_i, \mathbf{x}_j)$. Here, we set d as the product of the size of patch-pair $\mathbf{x}'_i$ and one constant factor, which is fixed to be 4 empirically, for the sake of both robustness and computational efficiency.

Let $P_{i,j}$ denote the co-occurrence frequency between patch-pairs $\mathbf{x}_i$ and and $\mathbf{x}_j$, we can extend Eq. (4) to impose our proposed co-occurrence prior, as follows:

$$\min_h \frac{\beta}{2} \sum_{i,j} (h_i - h_j)^2 P_{ij} + \alpha \sum_i h_i (1 - c_i), h \geq 0. \quad (5)$$

where α and β are tunable parameters. Herein, minimizing the first term, which relates the difference of two desired patch-pair confidences with their co-occurrence frequency, enforces that two frequently co-occurred patch-pairs are likely to bear similar semantic SCCs, while combining the first and the second terms into one objective function can further guarantee that those frequently co-occurred intr-class patch-pairs possess high SCCs at the same time.

We also verify the statistical observation of patch-pair co-occurrence prior on the MSRC [18] database. Every patch-pair of the given image is described by a 59-dimensional LBP descriptor extracted from the image regions covered by these two patches. We first retrieve for each patch-pair 1000 visually similar patch-pairs from the unlabeled image corpus based on Euclidean distance, and then compute the co-occurrence frequencies of every couple of patch-pairs, using the spatial histogram. We normalize each co-occurrence frequency by the maximum value within the same image, and plot the normalized frequencies in descending order. Figure 5 shows the comparison between the couples of intra-class patch-pairs (red color) and the couples of inter-class patch-pairs (blue color) from 10 images containing different semantic concepts. The horizontal axis represents the indices of patch-pair couples and the vertical axis represents their normalized co-occurrence frequencies. We can observe that the intra-class patch-pair couples usually have higher co-occurrence frequencies as compared to their counterparts, which well validates the proposed second-order patch-pair prior.

3.5. Unified Formulation

For each patch-pair of the given image, our intermediate goal is to augment its feature-based affinity into a seman-

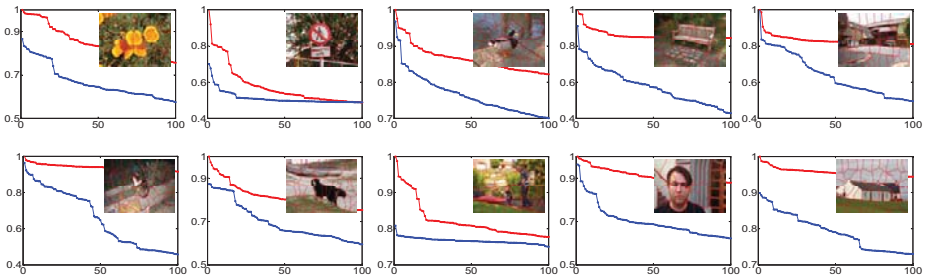

Figure 5. Co-occurrence frequency comparison between the couples of intra-class patch-pairs (red color) and the couples of inter-class patch-pairs (blue color). The images are from MSRC database [18]. The horizontal axis means the indices of patch-pair couples and the vertical axis means their normalized co-occurrence frequencies.

tic consistence confidence by harnessing both the first-order and the second-order patch-pairs priors. Integrating the Eqs. (2) and (5), we define a unified objective function to optimize the SCCs as follows,

$$\min_h L(h) = \frac{\beta}{2}\sum_{i,j}(h_i - h_j)^2 P_{ij} + \alpha\sum_i h_i(1-c_i) + \frac{1}{2}\sum_i (h_i - w_i)^2, \quad s.t. \quad h \geq 0, \tag{6}$$

where $h, c, w \in R^n$. The goal of the above objective function is to optimize the semantic consistence confidence for every patch-pair with in the input image. It is worthy noting that, since the intrinsic patch-pair affinity w_i is generally more informative than the corresponding patch-pair density c_i, we usually use a looser/smaller regularization parameter α.

We can further rewrite Eq. (6) as

$$\min_h L(h) = \frac{\beta}{2}Tr(h^T L h) + \alpha h^T(\mathbf{1} - c) + \frac{1}{2}\|h - w\|^2, \quad s.t. \quad h \geq 0, \tag{7}$$

where $L = D - P, D = diag\{\ldots, \sum_j P_{ij}, \ldots\}$ and L indicates the Laplacian matrix of matrix $P \in R^{n\times n}$. The objective function defined in Eq.(7) is quadratic and convex with respect to h, but there does not exist a closed-form solution due to the nonnegative constraint. In the next section, we shall develop an efficient iterative procedure which is scalable to the case with large number of patch-pairs.

3.6. Ultimate Image Segmentation

Suppose the patch-pair confidences are derived from Eq. (7) for one given image, our ultimate solution to patch grouping starts with constructing a similarity graph, each vertex of which represents one atomic patch. Two vertices are connected if the corresponding confidence h_i is greater than zero or a certain threshold, and the edge is weighted by h_i. The task of patch grouping can then be cast into finding a partition of the graph such that the edges between different groups have low weights while edges within a group have high weights. This problem can be solved by employing various spectral clustering techniques. In this work, we choose the popular Normalized Cut method [17], owning to its advantages in both efficiency and robustness against noises.

4. Multiplicative Iterative Solution

As aforementioned, we utilize iterative procedure, instead of general constrained quadratic optimization solver, for optimizing the problem (7).

Let ψ_i be the Lagrange multiplier for constraints $h_i \geq 0$, and $\psi = [\psi_i]$, the Lagrange $\mathcal{L}$ is then

$$\mathcal{L}(h) = \frac{\beta}{2}Tr(h^T L h) + \alpha h^T(\mathbf{1} - c) + \frac{1}{2}\|h - w\|^2 + Tr(\psi h^T) \tag{8}$$

Thus the partial derivative of $\mathcal{L}$ with respect to h is

$$\frac{\partial \mathcal{L}}{\partial h} = \beta L h + \alpha(\mathbf{1} - c) + (h - w) + \psi. \tag{9}$$

Along with the Karush-Kuhn-Tucker (KKT) condition [8] of $\psi_i h_i = 0$, we get the following equation,

$$\beta(Lh)_i h_i + \alpha(\mathbf{1} - c)_i h_i + (h_i - w_i)h_i = 0. \tag{10}$$

which leads to the following multiplicative update rule:

$$h_i \leftarrow h_i \frac{w_i}{(\beta L h + \alpha(\mathbf{1} - c) + h)_i}. \tag{11}$$

As the objective function is convex, the update rule Eq. (11) will converge to the global minimum of the problem (7). Related proofs can be referred to the past literature [9], or the general solution to the nonnegative second-order optimization problems proposed by Liu *et al*. [12].

5. Experiments

In this section, we evaluate the effectiveness of the discovered patch-pair priors and the proposed unified formulation for semantic image segmentation on several publicly available databases.

5.1. Databases

We use two publicly available databases, namely, MSRC [18] and COREL CDs [22] [1]. The MSRC database contains 590 images from 23 categories/labels with region-level ground-truths. The other database is from the widely used COREL collection and we use the subset provided in [22], which includes 800 images from 11 labels and provides the corresponding region-level ground-truth. We evenly partition each database into a testing set and another set used as the unlabeled image corpus. We resize all images with a ratio of $400/max(width, height)$. In addition, we use the 23 labels provided in MSRC database as query keywords on Microsoft BING image search engine to assemble an additional unlabeled image corpus. For each label/keyword, the top 100 online images returned by BING are downloaded. Thus, we obtain a corpus of 2300 images in total. We partition each image into a set of atomic patches using the method introduced in Section 2.1 and combine every two patches within the same image to form the patch-pair set. Each patch or patch-pair is described by a 59-dimensional Local Binary Pattern (LBP) descriptor [1]. Note that we choose to use one single feature type, rather than the combination of various features which has shown to be more effective for image problems in the past literature, because feature extraction, although important, is not the studying focus of this work.

5.2. Baselines and metrics

We adopt the widely used Normalized Cuts (NCut) [17] to perform patch clustering on the similarity graphs obtained by four different methods. The first one is to directly calculate the appearance similarity between every two different patches based on Gaussian similarity function. We denote this original affinity graph as *G-I* for ease of representation. The second graph is based on the semantic consistence confidence values from Eq. (7). We denote this graph as *G-II*. In order to evaluate how individual components, i.e. patch-pair density prior or patch-pair co-occurrence prior, take effects on the ultimate segmentation results, we set the tuning parameters α in Eq. (7) to be 0 and solve the equation to construct the patch affinity graph *G-III*, without considering the patch-pair density prior. Similarly, we set the parameter β to be 0 to obtain another patch affinity graph *G-IV*, without considering the patch-pair co-occurrence prior. Thus, we compare four algorithms, including NCut+G-I, NCut+G-II, NCut+G-III and NCut+G-IV, all of which work on the over-segmented images. We also use the patch-pair affinities calculated by above four algorithms, to directly predict whether two patches belong to the same semantic region, and compute the ROC curves for each affinity graph. Such a comparison can provide an overall picture at how the graphs obtained by different methods contribute to patch-pair classification.

Moreover, we implement three popular segmentation algorithms, namely, Multi-scale Normalized Ncut (MNCut) [17], Mean Shift [4] and the Graph-based segmentation proposed by Felzenszwalb *et al*. in [6]. All above three algorithms directly work on the original images. Note that in this work, we focus unsupervised scenarios and hence do not compare our method with the supervised segmentation algorithms.

We perform various image segmentation algorithms under a variety of scale parameters and report the best result. This evaluation strategy is also used in previous works, e.g. [2]. For NCut based algorithms, the number of segments is set within $\{2, 3, \ldots, 15\}$. The minimum region size of MeanShift algorithm is set within $\{1500, 2000, \ldots, 5000\}$ pixels. In implementation of our approach, the regularization parameters β and α are fixed at 0.5 and 0.2 respectively.

Two popular metrics [2] are used in this work for measuring the image segmentation performance. 1) *Segment covering rate* (**CR**). We define the covering rate of one segmentation S by the other segmentation S' as $\mathcal{C}(S' -> S) = \frac{1}{N}\sum_{R\in S}|R|\max_{R'\in S'}\mathcal{O}(R, R')$, where N is the number of segments within S and $\mathcal{O}(R, R') = \frac{|R \bigcap R'|}{|R \bigcup R'|}$. Herein R indicates one semantic region. This metric indicates how well the segmentation S can be explained by another segmentation S'. We evaluate different segmentation algorithms by calculating the covering of the ground truth segmentation by the derived segmentation. 2) *Variation of information* (**VI**), which measures the distance between two segmentations in terms of their average conditional entropy given by $VI(C, C') = H(C) + H(C') - 2I(C, C')$, where H and I represent respectively the entropies and mutual information between two clusterings of data C and C'. Smaller value of VI indicates better performance.

5.2.1 Results and analysis

Qualitative evaluations: We evaluate all the algorithms on an Intel Xeon $X5450$ workstation with 3.0 GHz CPU and 16 GB memory. The algorithms are implemented on MATLAB platform. Generally, it takes about 10 seconds to process one image for our proposed algorithm given that the features of atomic patches or patch-pairs have been extracted offline. We show some exemplar comparison results of semantic image segmentation in Figures 6 and 7. The images are from the testing subsets of MSRC and COREL

[1] We did not use the popular PASCAL dataset, since it is designed for given object segmentation (for the given 20 categories and supervised) and the labeled background parts still include many other object categories beyond the given 20 categories, thus our proposed segmentation algorithm shall further segment the background category into many regions and the provided ground-truths do not make sense in our scenario. We also did not use the Berkeley dataset, because most of the segments in groundtruth are not specified to one certain label, which is not applicable for our proposed approach.

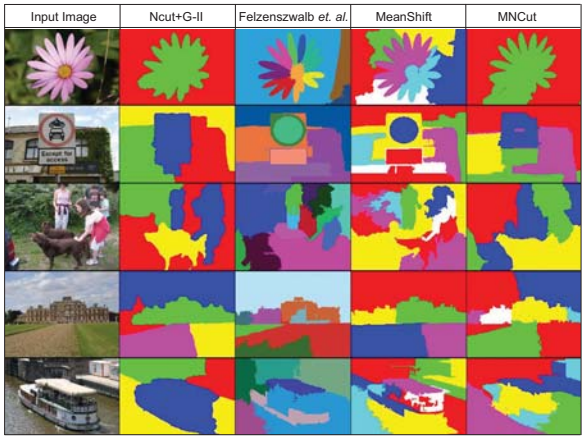

Figure 6. Exemplar comparison of segmentation results by various approaches on MSRC database [18].

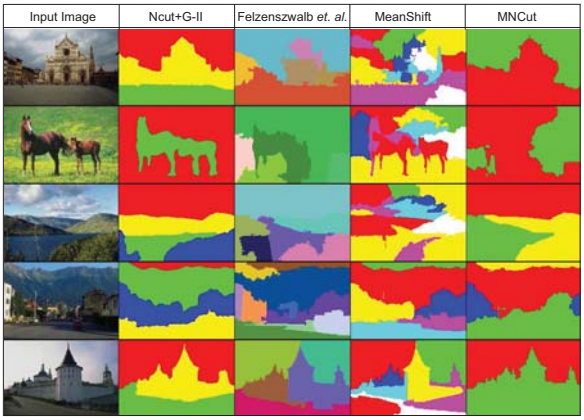

Figure 7. Exemplar comparison of segmentation results by various approaches on COREL database [22].

databases, respectively. In the figures, each row shows the original image in column 1 and the corresponding segmentation results obtained by algorithm Ncut+G-II, Felzenszwalb *et al*. [6], MeanShift [4] and MNcut [17] in columns 2, 3, 4 and 5, respectively. Different semantic regions are indicated with different colors. For each algorithm, the best segmentation results in terms of CR metric are reported while exploring different scale parameters. For algorithm Ncut+G-II, the unlabeled image corpus used here is from the BING Search Engine. These comparison under various conditions well validate the effectiveness of our approach.

Quantitative Evaluations: We report in Table 1 the accuracy comparison among various image segmentation algorithms on both MSRC and COREL databases. Furthermore, Figure 8 shows the ROC curves obtained by directly applying the affinity values in G-I, G-II, G-III, and G-IV for patch classification. The horizontal axes indicates the false positive rates and the vertical axes indicates the true positive rates. Herein, G-II, G-III and G-IV are computed by using the unlabeled image corpus from the BING search

Table 1. Performance comparison of various algorithms on MSRC and COREL databases. The best results achieved are indicated with bold font in each column.

	MSRC		Corel CDs	
Metrics	CR (%)	VI	CR (%)	VI
Mean Shift [4]	41.90	2.50	50.69	1.92
Felzenszwalb *et al*. [6]	51.02	2.08	60.94	1.61
MNCut [17]	59.22	1.66	68.73	1.24
NCut+G-I [17]	58.33	1.60	68.08	1.20
Unlabeled Image Corpus from the Evaluation Databases				
NCut+G-II	64.66	1.47	75.03	1.03
NCut+G-III	61.07	1.54	70.53	1.15
NCut+G-IV	61.45	1.53	72.13	1.07
Unlabeled Image Corpus from the BING Search Engine				
NCut+G-II	**65.24**	**1.45**	**75.28**	**1.03**
NCut+G-III	61.32	1.55	70.70	1.14
NCut+G-IV	61.55	1.54	72.01	1.08

engine. From these results, we can obtain the following observations. 1) The proposed solution, namely NCUT+G-II, achieves much better performances on both databases as compared to other baselines. For example, the covering rates of NCut+G-II on the MSRC and COREL databases are 65.24% and 75.28%, which outperform NCut+G-I with the margins of 6.91 and 7.20 percentages, respectively. Similar improvements can also be achieved in terms of VI metric. This clearly demonstrates the effectiveness of the proposed approach which employs the auxiliary patch-pair priors for semantically patch grouping. 2) The segmentation results achieved by the algorithm NCut+G-II using the online images as unlabeled image corpus are slightly better than that using the images from the same database. This is due to the fact that larger image corpus usually provides more robust statistical information for our proposed patch-pair priors. It is worthy noting that although the search results returned by the BING search engine contain large variations, our approach can still achieve high-quality image segmentation. 3) Algorithm NCut+G-II achieves better performance than algorithms NCut+G-III and NCut+G-IV while the latter two algorithms achieve better performances than other four baselines, on both databases. These comparisons well justify the effectiveness of the proposed patch-pair priors for semantic image segmentation.

In addition, it is worthy highlighting following discussions. 1) The newly discovered priors can be used, but not limited to, in the formulation of this work to help improve segmentation accuracies. Also, it is easy to combine our discovered priors with any other state-of-the-art models or algorithms to obtain better segmentation performance. 2) We did not try to combine the edge/boundary information, which has been widely used in past efforts on

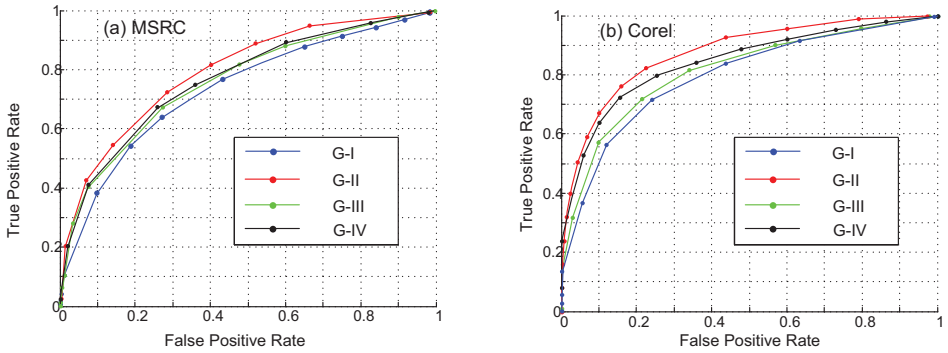

Figure 8. ROC curves. We apply various affinity graphs for directly patch classification on (a) MSRC database [18] and (b) COREL database [22].

image segmentation problem, because this work focuses on region/patch grouping problem. However, our proposed priors and formulation can be easily extended to integrate such important information for further boosting segmentation accuracy.

6. Conclusions

In this work, we reported two discriminative priors, i.e., patch-pair density prior and patch-pair co-occurrence prior, and showed how these priors from unlabeled image corpus can boost semantic image segmentation. Also, a unified formulation was developed to combine the discovered patch-pair priors mined from the auxiliary unlabeled image corpus and the intrinsic feature-based patch-pair affinities extracted from the given image itself. The objective function is optimized by efficiently and scalable multiplicative updating rules. Significant improvements on image segmentation accuracies were achieved on two image evaluation sets. In contrast with previous efforts, our work suggests two simple yet informative patch-pair priors and a unified formulation, both of which can be used for other kinds of image applications, such as part-based object recognitions and image retrieval.

7. Acknowledgement

This research is done for CSIDM Project No. CSIDM-200803 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore. Also it is partially supported by the China 863 program (Grant No.$2008AA01Z126$) and the National Natural Science Foundation of China (under Grant No. 60970156).

References

[1] T. Ahonen, A. Hadid, and M. Pietik$\grave{a}\dot{i}$nen. Face description with local binary patterns: application to face recognition. *TPAMI*, 2006.

[2] P. Arbel$\acute{a}$ez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *CVPR*, 2009.

[3] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color and texture-based image segmentation using em and its application to content-based image retrieval. In *ICCV*, 1998.

[4] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *TPAMI*, 2002.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* John Wiley & Sons, New York, 2 edition, 2001.

[6] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[7] J. Hays and A. Efros. Scene completion using millions of photographs. *ACM TOG*, 2007.

[8] H. Kuhn and A. Tucker. Nonlinear programming. In *2nd Berkeley Symposium*, 1951.

[9] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

[10] S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum. Statistical learning of multi-view face detection. In *ECCV*, 2002.

[11] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[12] X. Liu, S. Yan, J. Yan, and H. Jin. Unified solution to nonnegative data factorization problems. In *ICDM*, 2009.

[13] X. Liu, J. Feng, S. Yan and H. Jin. Image Segmentation with Patch-Pair Density Priors. In *ACM MM*, 2010.

[14] G. Mori. Guiding model search using segmentation. In *ICCV*,2005.

[15] B. Russell, A. Efros, J. William, F. Freemand, and A. Zisserman. Segmenting scenes by matching images composites. In *NIPS*, 2009.

[16] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *ICCV*, 2003.

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for mulit-class object recognition and segmentation. In *ECCV*, 2006.

[19] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *TPAMI*, 2008.

[20] Z. Tu and S. Zhu. Image segmentation by data-driven markov chain monte carlo. *TPAMI*, 2002.

[21] A. Yang, J. Wright, Y. Ma, and S. Sastry. Unsupervised segmentation of natural images via lossy data compression. *CVIU*, 2008.

[22] J. Yuan, J. Li, and B. Zhang. Exploiting spatial context constraints for automatic image region annotation. In *ACM MM*, 2007.