Robust Region Grouping via Internal Patch Statistics

Xiaobai Liu[1], Liang Lin[2], Alan L. Yuille[1]
[1]Department of Statistics, University of California at Los Angeles, CA 90095
[2]Sun Yat-Sen University, Guangzhou 510006, China
lxb@ucla.edu, linliang@ieee.org, yuille@stat.ucla.edu

Abstract

In this work, we present an efficient multi-scale low-rank representation for image segmentation. Our method begins with partitioning the input images into a set of superpixels, followed by seeking the optimal superpixel-pair affinity matrix, both of which are performed at multiple scales of the input images. Since low-level superpixel features are usually corrupted by image noises, we propose to infer the low-rank refined affinity matrix. The inference is guided by two observations on natural images. First, looking into a single image, local small-size image patterns tend to recur frequently within the same semantic region, but may not appear in semantically different regions. We call this internal image statistics as replication prior*, and quantitatively justify it on real image databases. Second, the affinity matrices at different scales should be consistently solved, which leads to the cross-scale consistency constraint. We formulate these two purposes with one unified formulation and develop an efficient optimization procedure. Our experiments demonstrate the presented method can substantially improve segmentation accuracy.*

1. Introduction

Image segmentation is to partition input image into several semantically consistent regions. It is one of the most challenging and actively studied problems in computer vision. The difficulties are due to the large intra-category variations as well as the ill-posed nature of segmentation. In the past literature, many efforts have been devoted to extra information beyond the input image for solving this problem [17, 13]. Although impressive results and successes achieved, these methods have two major limitations. First, analyzing the statistics of a large number of related images itself is challenging because the low-level visual features are usually not powerful enough. Therefore, the quality of segmentation heavily depends on how well the extra images match with the given image [17]. Second, fully exploring the cross-image statistics is obviously time-consuming and computationally inefficient, which also limits the application of these methods.

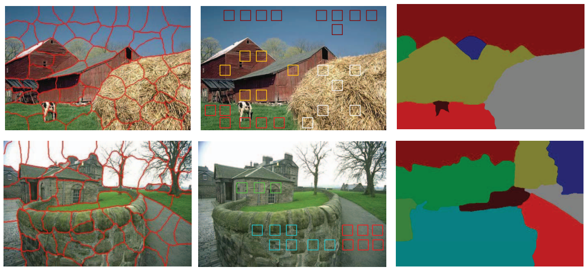

Figure 1. Illustrations of the discriminative internal image statistics. (a) two input images overlaid with superpixel over-segmentation results; (b) repeatedly occurred patches (identified with the same color); (c) segmentation results by our algorithm(unsupevised).

In this work, we introduce a simple yet efficient internal image statistics for image segmentation, and integrate it within a unified low-rank image representation. Our method is based on the multi-scale image representation. For each scale of the input image, we partition it into a set of non-overlapping superpixels [16], and construct an affinity graph by taking superpixels as graph vertices. As conventionally, each superpixel is represented as one appropriate appearance feature, e.g., color. Given these superpixel features, we assume that intra-class superpixels, namely the superpixels belonging to the same semantic region, are drawn from one identical low-rank feature subspace. Our goal is to seek for a low-rank refined superpixel affinity matrix, so that: the intra-class superpixel affinities are dense, whereas the inter-class superpixel affinities are all sparse or zeros. As shown in the previous work [14], low-rank constraint is helpful to suppressing the effects of data noises and corruptions.

The inference of low-rank refined affinity matrices is guided by two more purposes. The first one is a discriminative observation on natural images: *small-size image patches (e.g.* 6×6 *pixels) with certain appearance patterns tend to recur frequently within the same semantic region,*

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.252

but may not appear in semantically different regions. We call this internal statistics as *replication prior*.

Fig. 1 illustrates this observation in the middle column, where the colored rectangles indicate the small-size image patches extracted from the image grid. We can intuitively find that a small-size patch usually has multiple copies (identified with the same color) in the same semantic region (shown in the right column). The replication prior can be used to measure how likely two subregions [1] are semantically similar. Generally, if every patch within one subregion has many copies in another one, namely these two subregions have high replication prior, they will belong to the same semantic region with high probability, and vice versa. In later sections, we will further quantitatively justify that the above observation on fairly small-size patches is partially true in real image databases. Replication prior can serve as a discriminative cue for image segmentation. It is different yet complementary with the low-level features, e.g. color, gradient, that are directly extracted from image regions [16].

The second purpose lies on the fact that the desired superpixel-pair affinities at different scales should be consistently solved, which leads to the so-called cross-scale consistency constraint. We formulate the pursuit of low-rank refined affinity matrices and the above two purposes as a unified constraint nuclear norm and ℓ_1-norm minimization problem. The obtained formulation is a convex quadratic function and we introduce an efficient optimization procedure based on the augmented Lagrange multiplier (ALM) method [12]. Taking the solved low-ranked refined superpixel affinities, one can call the Normalized Cut method [6] to address the unsupervised segmentation problem.

The contributions of this work are two-fold. First, we develop a multi-scale low-rank representation to seek for the affinity matrix at each scale in parallel, while preserving the cross-scale consistency. Second, we study a simple yet efficient internal image statistics, and present a practical method for image segmentation. The advantages of our approach are demonstrated by extensive experiments with comparisons to the popular segmentation algorithms on two public datasets MSRC [19] and BSD500 [15].

1.1. Relationship to Previous Works

Our approach is closely related to the efforts in image segmentation for which a broad family of unsupervised methods have been proposed. The typical ones include Tu and Zhu's data driven MCMC algorithm [20], Comaniciu and Meer's Mean-shift [7], and Shi and Malik's Normalized Cuts method (NCut) and its extension multi-scale NCut [6]. All these three algorithms directly utilize the low-level feature descriptors which makes the segmentation task painfully unconstrained. In contrast, our method aims to utilize both the low-level features and the extra discriminative prior. As aforementioned, extra knowledge or statistics have been studied to address the ill-posed nature of image segmentation [13] [17]. Nevertheless, these algorithms usually requires hundreds or even thousands of images to achieve robust segmentation, which leads to the dilemma of balancing the matching accuracy and the computation cost. In contrast, the proposed replication prior is a kind of internal image statistics, which bears the obvious benefit of low computational demand. Moreover, we will experimentally show that, to achieve the equally good quality of segmentation by the low-rank refined replication prior, hundreds of images are required for the external statistics [13].

In computer vision literature, internal image statistics has been used in various low-level image tasks, e.g., texture synthesis [8], denoising [3], and super-resolution [10]. Bagon *et al*. [4] assume that one semantic region can be well explained by repeatable compositions and utilized this assumption for interactive image segmentation which requires user input. Recently, Zontak and Irani [23] further quantitatively evaluate the strength of internal statistics, and demonstrate its advantages in enhancing the quality of image denoising and super-resolution. Our work extends these methods, and introduces a practical method to apply the internal image statistic for image segmentation. The developed framework can be used for other high-level image tasks which will be explored in future.

Our method is also motivated by the advances in subspace clustering, especially the low-rank representation (LRR) method, which seeks for the lowest-rank representation among possible candidates that represent all samples as the linear combination of the bases in a dictionary. The bases are assumed to be sampled from either one single subspace, e.g., the matrix competition method [5], or a union of multiple linear subspaces [14]. The advantages of LRR in robustness have been well demonstrated in comparisons to other subspace clustering methods, e.g. sparse subspace clustering [21], especially while encountering with severe corruptions and noises. Our method extends the LRR method to infer multiple low-rank representations separately at multiple scales, and simultaneously preserve the cross-scale consistency.

2. Multi-scale Low-rank Image Representation

Given an image as input, we resize it with several factors and oversegment each scaled image into a set of non-overlapping super-pixels. We choose to use the method developed by Ren and Malik in [16]. Each superpixel comprises an ensemble of pixels that are spatially coherent and perceptually similar with respect to certain appearance fea-

[1] A *subregion* indicates a part of the whole image, either a semanticless superpixel or a semantic region (e.g., cars, buildings). A subregion usually contains multiple small-size patches.

tures (e.g. color). Let M denote the number of scales, I^s be the scaled image indexed by $s = 1, .., M$. We construct an affinity graph by taking the superpixels in I^s as graph vertices. Every two different vertices are connected by a graph edge, weighted by the likelihood of the associated superpixel-pair belonging to the same semantic region. Thus, the overall quality of segmentation depends on the pairwise superpixel affinity matrix. Let $x_i^s \in R^d$ denote the d-dimension feature descriptor extracted for the i^{th} superpixel in I^s. R denotes a real matrix. One common method to compute the superpixel-pair affinity [6] is $exp(-\|x_i^s - x_j^s\|^2)$, where $\|\cdot\|$ indicates the Frobenius norm of a vector or a matrix. Nevertheless, because of various image noises and clutters, the obtained affinity matrix is always corrupted and not discriminative enough to produce high-quality segmentations. We will introduce in the rest of this section a novel formulation to infer more powerful superpixel-pair affinities for the input image, and further discuss how to apply this representation to segmentation problems in next section.

2.1. Objective-I: Low-rank Image Representation

The major step of our method is to pursue the low-rank refined affinity matrix from the low-level superpixel features. Herein, we assume that superpixels belonging to the same semantic region are all drawn from the same low-rank feature subspace, and all superpixels in the same image (also the same scale) lying on a union of multiple subspaces [14]. We aim to represent each superpixel descriptor as a linear combination of other superpixel descriptors, and seek for the lowest rank representation of all superpixels in a joint fashion. Let n^s denote the number of superpixels in I^s, and $X^s = [x_1^s, x_2^s, \ldots, x_{n^s}^s] \in R^{d \times n^s}$. Each vector x_i^s can be represented as the linear combination of the column vectors of X^s, denoted as,

$$x_i^s = X^s z_i^s, \tag{1}$$

where $z_i^s \in R^{n^s}$ is the coefficient vector. Large z_{ij}^s generally indicates that x_i^s and x_j^s have similar projection in the feature subspaces spanned by the column vectors of X^s, and vice versa. Thus, we can use z_{ij}^s to measure the affinity between superpixels i and j. Denote $Z^s = [z_1^s, z_2^s, ..., z_{n^s}^s]$ as the desired affinity matrix. Since the data vectors in X^s are often noisy or grossly corrupted, we relax Eq. (1) to allow a fraction of X^s are corrupted. Padding the coefficient vectors into matrix form, we have,

$$X^s = X^s Z^s + E^s, \tag{2}$$

where $E^s \in R^{d \times n^s}$ denotes the error matrix. The lowest rank representation of the superpixel affinity matrix Z^s can be solved by following program

$$\begin{aligned} \min_{\{Z^s\},\{E^s\}} & \sum_{s=1}^{M} \|Z^s\|_* + \lambda \|E^s\|_1 \\ s.t. \quad & X^s = X^s Z^s + E^s, \end{aligned} \tag{3}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, $\|\cdot\|_1$ denotes the ℓ_1-norm of a matrix, and the parameter $\lambda > 0$ is used to balance the effects of the two parts. The ℓ_1-norm in Eq. (3) is used to enforce most of the elements of E^s have zero errors, namely, only partial data vectors are corrupted and others are clean.

2.2. Objective-II: Replication Prior

The discovered replication prior is from a statistical observation on natural images: *local small-size patches (e.g. 6×6 pixels) tend to recur frequently within the same semantic region, yet less frequently within semantically different regions*. The small-size patches are collected from the image grid that evenly partitions the image into non-overlapping rectangles. Fig. 1 shows the patches in the middle column. The size of image patches is fairly small so one superpixel may contain multiple patches. Replication prior can be used to measure the semantic consistency of two superpixels.

We use patch recurrence density to quantify the replication prior. From each small-size patch, indexed as p, we extract a feature descriptor, denoted as f_p. Let Λ denote a subregion and q index the patches in Λ. The empirical density of p with respect to Λ can be estimated using the Parzen window method:

$$D(p, \Lambda) = \frac{1}{|\Lambda|} \sum_{q \in \Lambda} \delta(\mathcal{K}(\|f_p - f_q\|^2), \zeta) \tag{4}$$

where $\mathcal{K}$ is a Gaussian kernel, and $|\Lambda|$ denotes the number of patches in Λ. $\delta(\cdot, \zeta)$ returns the first parameter if it is larger than the constant ζ, or 0 otherwise. Since Parzen method does not distinguish between the smaller number of perfectly similar patches, and the larger number of partially similar patches, we introduce a constant threshold ζ to depress the effects of partially similar patches.

We perform an experiment on the Berkeley Segmentation Database [15] to justify the replication prior. We use the test subset of 200 images. Only the original scale (about 480 × 320 pixels) is used. For each patch p, we extract a 31-dimension histogram of oriented gradient (HOG) [9], and compute its intra-class density and inter-class density. The intra-class density of patch p is calculated using Eq. (4) where Λ is set to be the semantic region in groundtruth that contains the patch p. Correspondingly, the inter-class density of patch p is estimated with respect to the semantic region that does not contain patch p. Since there usually exists

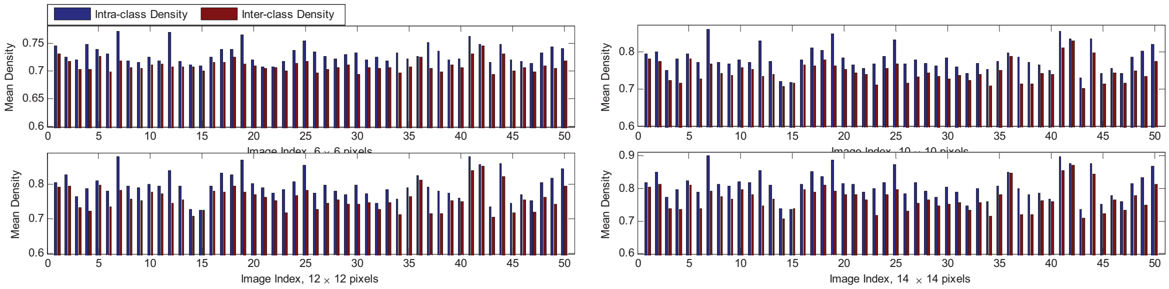

Figure 2. Discriminative Replication Prior. We compare the mean intra-class patch densities and mean inter-class patch densities under different patch sizes, including 6 × 6, 10 × 10, 12 × 12, and 14 × 14 pixels. The images are from BSD database [15]. See texts for more details.

more than one semantically different regions for patch p, we select the highest one (or the most ambiguous one) as the inter-class density of patch p . Next, we compute for each image the mean intra-class density and the mean inter-class density by averaging over all patches. Figures 2 plots the density comparisons while looking at different patch-sizes, including 6×6, 10×10, 12×12 and 14×14 pixels, respectively. The horizontal direction indicates the image indices, and the vertical direction indicates the mean patch densities. Herein, we only plot the densities of 50 images due to the space limitation. We can observe that intra-class densities are usually higher than inter-class densities under different patch-sizes. Similar observations can also be reached on the rest 150 images, while looking at different scales. This experiment shows that the discovered replication prior is partially true while using fairly small-size patches.

We utilize the replication prior to measure how likely two superpixels belong to the same semantic region. Let Λ^s_i denote the subregion in I^s covered by superpixel i, Q^s denote a $n^s \times n^s$ matrix. We have,

$$Q^s_{ij} = e^{-(\frac{1}{|\Lambda^s_i|}\sum_{p\in\Lambda^s_i} D(p,\Lambda^s_j)+\frac{1}{|\Lambda^s_j|}\sum_{q\in\Lambda^s_j} D(q,\Lambda^s_i))}. \qquad (5)$$

Large Q^s_{ij} indicates the associated suerpixel-pair has low replication prior, namely the superpixels i and j belong to different semantic regions with high probability, and vice versa. Once estimated Q^s, we can use it to regularize the inference of the desired low-rank representations in Eq. (3) through following program,

$$\min_{\{Z^s\},\{E^s\}} \sum_s \|Z^s\|_* + \lambda\|E^s\|_1 + \beta tr({Z^s}^T Q^s)$$
$$s.t. \quad X^s = X^s Z + E^s \qquad (6)$$

where ${Z^s}^T$ indicates the transform of Z^s, $tr(\cdot)$ returns the matrix trace, and β is a constant tuning parameter. Minimizing the term $tr({Z^s}^T Q^s)$ will encourage the intra-class superpixel-pairs have higher affinities (in Z^s) than the inter-class superpixel-pairs. In this way, replication prior is used as a kind of soft constraint to regularize the inference of the lowest rank affinity matrices from the low-level visual features. Eq. (6) provides a general way to fuse two different yet complementary representations of the same data, which can be used for other tasks. It is worthy noticing that the replication prior is based on a high-level semantic statistics, which, although estimated using low-level appearance features, is able to describe how the small-size patches form the large-size semantic regions and thus provides more knowledge of structure about the solution space to image segmentation.

2.3. Objective-III: Cross-scale Consistency

The objective function in Eq. (6) aims to compute the optimal superpixel affinity matrix for each scale separately. But it is crucial to propagate knowledge from one scale to another scale which has been justified in the past works on multi-level image methods, e.g., [6]. We can achieve this by projecting the superpixels at the coarse-level to the fine-level and introducing a cross-scale consistency constraint.

Formally, let I^s_i indicate the superpixel at the scale s indexed by i. We project I^s_i to the coarse-scale $s+1$, and if one super-pixel I^{s+1}_j overlaps with I^s_i more than half the area of I^{s+1}_j, we call I^{s+1}_j as the parent of I^s_i. We impose a cross-scale constraint for every two neighbor scales (namely the coarse-scale $s+1$ and the fine-scale s): *for every two superpixels at the coarse-level, their affinity should be locally average of the affinities between their respective children at the fine-level.* We arrange all elements of Z^s in one column vector

$$f(Z^s) = \left[\begin{array}{ccccc} Z^s_{11} & Z^s_{12} & .. & Z^s_{21} & .. \end{array}\right]^T \in R^{(n^s\times n^s)\times 1}. \qquad (7)$$

Let $h^{s+1}(I^s_a)$ indicate the index of the superpixel in I^{s+1} which is the parent of I^s_a. Let $k =< a, b >$ index the elements or superpixel-pairs of $f(Z^s)$ that formed by I^s_a and I^s_b, and $l =< c, d >$ index the elements of $f(Z^{s+1})$. For every two neighbor scales, we first construct a constant matrix, denoted as $C^{s+1,s}$, as follows.

i). Set $C^{s+1,s} \in R^{(n^{s+1}\times n^{s+1})\times(n^s\times n^s)} = 0$;

ii). For $\forall k =< a, b >, k = 1, .., n^s; \forall l =< c, d >, l = 1, .., n^{s+1}$,

if $h^{s+1}(I^s_a) == h^{s+1}(I^s_b)$,

set $C^{s+1,s}(l, k) = 0$;

else

if $(h^{s+1}(I^s_a) == c, h^{s+1}(I^s_b) == d)$ or $(h^{s+1}(I^s_a) == d, h^{s+1}(I^s_b) == c)$

set $C^{s+1,s}(l, k) = 1$;

iii). Normalize $C^{s+1,s}$ so that the summarization of each row is 1;

Then, the consistency constraint between the coarse-scale $s + 1$ and the fine-scale s can be encoded by minimizing a least square function,

$$\min_{Z^s, Z^{s+1}} \|C^{s+1,s} f(Z^s) - f(Z^{s+1})\|^2 \tag{8}$$

Furthermore, let $Z = [f(Z^1); f(Z^2), \dots, f(Z^M)]$, and

$$C = \begin{bmatrix} C^{2,1} & -\mathbf{1}^2 & & 0 \\ & ... & ... & \\ 0 & & C^{M,M-1} & -\mathbf{1}^M \end{bmatrix}. \tag{9}$$

where $\mathbf{1}^s$ indicates an identity matrix. We can obtain a compact form of the cross-scale consistency constraint,

$$\min_{\{Z^s\}} \|CZ\|^2 \tag{10}$$

Minimizing the above term will enforce the desired superpixel affinity matrices at different scales are consistent.

2.4. Unified Formulation

Integrating Eq. (6) and Eq. (10), we can define a unified formulation as,

$$\min_{\{Z^s\},\{E^s\}} \sum_{s=1}^{M} \|Z^s\|_* + \lambda\|E^s\|_1 + \beta tr({Z^s}^T Q^s) + \gamma\|CZ\|^2$$
$$s.t. \quad X^s = X^s Z^s + E^s \tag{11}$$

where γ is a tuning parameter.

2.5. Optimization

To optimize Eq. (11), we adopt the Augmented Lagrange Multiplier (ALM) method [12] due to its efficiency in convergence. By introducing the auxiliary variables $\{J^s\}, \{P^s\}, s = 1, .., M$, we convert Eq. (11) to an equivalent problem,

$$\min \sum_{s=1}^{M} \|J^s\|_* + \lambda\|E^s\|_1 + \beta tr({P^s}^T Q^s) + \gamma\|CZ\|^2$$
$$s.t. \; X^s = X^s P^s + E^s, \; Z^s = J^s, \; Z^s = P^s \tag{12}$$

Algorithm 1 Optimization Procedure to Eq. (12)

```
Input: data matrix $\{X^s\}, s = 1, .., M$, parameter $\lambda$, $\beta$ and $\gamma$;
Set $J^s = Z^s = P^s = E^s = 0, Y^s = 0, V^s = 0, U^s = 0, \mu = 10^{-6}; \varepsilon = 10^{-8}; \rho = 1.1$;
While not converged do (iteration body),
    1 Update $J^s$ by solving
        $J^s = \arg\min_{J^s} \frac{1}{\mu}\|J^s\|_* + \frac{1}{2}\|J^s - (Z^s + V^s/\mu)\|^2$
    2 Update $P^s$ by $P^s = ({X^s}^T X^s + \mathbf{1})^{-1}[{X^s}^T X^s - {X^s}^T E^s + Z^s + ({X^s}^T Y^s - \beta Q^s + U^s)/\mu]$;
    3 Update $Z$ by $Z = (\frac{\gamma}{\mu}C^T C + \mathbf{1})^{-1}B/2$, where the vector $B$ has following structure:
        $B = [\; f(G^1)^T \quad f(G^2)^T \quad ... \quad f(G^M)^T \;]^T$
        with $G^s = J^s + P^s - (V^s + U^s)/\mu, \;\; s = 1, .., M$
    4 Update $\{E^s\}$ by solving
        $E^s = \arg\min_{E^s} \frac{\lambda}{\mu}\|E^s\|_1 + \frac{1}{2}\|E^s - (X^s - X^s P^s + Y^s/\mu)\|^2$
    5 Update the multipliers and parameter as:
        $Y^s = Y^s + \mu(X^s - X^s P^s - E^s)$;
        $V^s = V^s + \mu(Z^s - J^s)$;
        $U^s = U^s + \mu(Z^s - P^s); \mu = \rho\mu$;
    6 Check the convergence condition: $\forall$s, $\|X^s - X^s Z^s - E^s\| < \varepsilon$;
```

The related unconstrained problem of (12) is defined as,

$$\min \gamma\|CZ\|^2 + \sum_s \|J^s\|_* + \lambda\|E^s\|_1 + \beta tr({P^s}^T Q^s)$$
$$+ < Y^s, X^s - X^s P^s - E^s > + \frac{\mu}{2}\|X^s - X^s P^s - E^s\|^2$$
$$+ < V^s, Z^s - J^s > + \frac{\mu}{2}\|Z^s - J^s\|^2$$
$$+ < U^s, Z^s - P^s > + \frac{\mu}{2}\|Z^s - P^s\|^2, \tag{13}$$

where $\mu > 0$ is the penalty parameter, $\{Y^s\}, \{V^s\}$ and $\{U^s\}$ are augmented Lagrange multipliers.

The ALM algorithm alternately solves the objective function in Eq. (12) w.r.t. $\{Z^s\}, \{E^s\}, \{J^s\}$, or $\{P^s\}$, with other variables fixed. All the subproblems have closed-form solutions, and the entire ALM procedure will converge Q-linearly, as shown in [12], while μ increasing. Algorithm 1 summarizes the optimization procedure. Note that: i) the optimal solution to the objective functions in Step-1 and Step-4 can be solved by the soft-threshold (or shrinkage) method in [12]; ii) the subproblems in Step-3 can be solved analytically; and iii) the update factor ρ is set to be 1.1 to obtain a series of increasing μ.

3. Implementation

We apply the proposed low-rank refined image representation for unsupervised image segmentation.

Suppose the optimal superpixel affinity matrices are solved from Eq. (11), denoted as $Z^{s*}, s = 1, .., M$. We first project Z^{s*} to the pixel-level as follows: if two pixels in I^s belong to the same superpixel, we set their affinity to be 1; otherwise, we set their affinity to be the corresponding superpixel-pair affinity. We define the affinity between neighboring pixels at scale s using the linear combination of a set of Gaussian kernels:

$$W^s_{ij} = \sum_{k=1}^{3} w^k \mathcal{K}^k_{ij} \tag{14}$$

where $\mathcal{K}^k$ is a Gaussian kernel in a specific feature space, w^k is linear combination weight. Here, i, j index the pixels, not superpixels, in images. In particular, $\mathcal{K}^1$is an appearance kernel, defined as $\mathcal{K}^1_{ij} = exp(-\alpha^1\Delta^a_{ij} - \alpha^2\Delta^l_{ij})$. Δ^a_{ij} indicates the Euclidean distance between the color vectors at pixel i and pixel j. Δ^l_{ij} indicates the geometrical distance between pixel i and j. $\mathcal{K}^2$ is the smoothness kernel, defined as $\mathcal{K}^2_{ij} = exp(-\alpha^3\Delta^l_{ij})$. These two kernels, following previous works [6, 19, 11], are used to represent the appearance and smoothness relationships between neighboring pixels. The third kernel is to utilize the repetition prior and takes the form of: $\mathcal{K}^3_{ij} = exp(-\alpha^4\Delta^r_{ij})$, where $\Delta^r_{ij} = exp(-{Z^{s*}}_{ij}/2\sigma^2)$. The settings of the parameters $\alpha^1, \alpha^2, \alpha^3, \alpha^4, w^1, w^2, w^3, \sigma$ are described in Section 4. We apply the multi-scale NCut (MNCut) method [6] on $W^s_{ij}, s = 1, 2, ...$, to achieve the final segmentation.

4. Experiments

In this section, we apply the discovered replication prior and the proposed multi-scale low-rank representation (MsLRR) for image segmentation and evaluate them on publicly available image databases.

4.1. Evaluation Settings

We use two databases, MSRC [19] and Berkeley Segmentation Database (BSD) [15]. MSRC contains 591 images.We divide the dataset into two equal-size subsets and use one for testing. BSD contains 500 images among which 200 images are used for test. There are five groundtruths per image, and we use the first one for evaluation. All images are resized with three factors 0.8, 1.0, and 1.5, and over-segmented into superpixels using the method in [16]. The number of superpixels is set to be 80, 120 and 150, respectively. We extract for each superpixel three types of feature descriptors, including 12-dimension color histogram in each channel of RGB, 59-dimension Local Binary Pattern and 31-dimension Histogram of Oriented Gradient [19]. These descriptors are first normalized by their respective ℓ_2-norms and then concatenated to form one single descriptor vector (i.e. x^s_i).

We use two segmentation metrics, *Covering Rate* (**CR**), namely the percentage of per-pixel agreement between the obtained results w.r.t the groundtruth, and *Variation of information* (**VI**) [1]. Smaller value of VI indicates better performance.

4.2. Exp-I: Unsupervised Image Segmentation

We apply the proposed approach to segment a given image. For comparisons, we implement four variants of the proposed method. 1) *MsLRR-I*, that sets $\beta = 0$ and $\gamma = 0$ in Algorithm 1. It degenerates to seeking for the low-rank refined affinity matrices at multiple scales separately. 2) *MsLRR-II*, that sets $\beta = 0$, and only uses the regularization term of cross-scale consistency constraint. 3) *MsLRR-III*, that sets $\gamma = 0$, and only uses the regularization term of replication prior . 4) *MsLRR-IV*, that utilizes both regularization terms. We also implement two popular image segmentations methods, the Multi-scale Normalized Cut method (MNCut) [6] and the Mean-Shift method [7]. Moreover, we compute the superpixel-pair affinities based on feature descriptors, namely letting $\Delta^r_{ij} = exp(-|x_i - x_j|^2/2\sigma^2)$ in Eq. (14) , and call MNCut to segment images. We refer to this algorithm as *MNCut-SP* and the original method [6] as *MNCut*. We perform various algorithms under a variety of scale parameters and report the best result. This evaluation strategy is also used in previous works [2] [18]. For MsLRR and MNCut based algorithms, the number of segments is set within {2, 3, . . . , 15}. The minimum region area of MeanShift algorithm is set within {1500, 2000, . . . , 10000} pixels. For MsLRR algorithms, we set $\beta = 0.001$, and $\gamma = 0.001$ if the related regularization term is used. Set $\lambda = 0.0001$. The patch-size is fixed to be 6 × 6 pixels. The threshold ζ for patch density estimation is fixed to be 0.4. Set w^1, w^2, w^3 as $0.4, 0.3, 0.3$ respectively. Set $\alpha^1, \alpha^2, \alpha^3, \alpha^4$ as $0.6, 0.4, 0.1, 0.1$ respectively, and $\sigma = 1$. We fix these parameters throughout the evaluations.

Table 1-(a) reports the performance comparisons of various unsupervised segmentation algorithms on MSRC [19] and BSD [15] databases. We also compare the available results of other state-of-the-art segmentation algorithms, that refer to as UCM [2], TBES [18] and CTM [22]. We can obtain following observations. 1) The proposed MsLRR algorithms can achieve comparable accuracies as the state-of-the-art methods UCM [2], and TBES [18], and much better results than other baseline methods. 2) MsLRR-IV clearly outperforms MsLRR-II and MsLRR-III, while the later two algorithms achieve higher accuracies than MsLRR-I on both databases. These comparisons well justify the effectiveness of MsLRR and the internal replication prior. 3) From the comparisons between MsLRR-II and MsLRR-I, one can see that the cross-scale consistency is helpful to achieving robust segmentation.

Table 1. Performance comparisons of unsupervised segmentation algorithms on MSRC [19] and BSD500 [15]

(a) Segment single image

	MSRC		BSD	
Metrics	CR (%)	VI	CR (%)	VI
MNCut [6]	60.64	1.59	52.31	2.18
UCM [2]	–	–	–	2.11
CTM [22]	–	–	-	2.02
MNCut-SP	62.94	1.48	57.97	1.91
Mean Shift [7]	62.37	1.54	58.81	1.83
TBES [18]	–	1.49	–	1.76
MsLRR-I	64.34	1.45	61.82	1.68
MsLRR-II	68.31	1.36	63.31	1.53
MsLRR-III	66.53	1.39	62.24	1.59
MsLRR-IV	**69.97**	**1.32**	**63.87**	**1.47**

(b) Using unlabeled images from the training subsets

MsLRR-V	68.78	1.34	62.57	1.56

(c) Using unlabeled images from the Internet

MsLRR-V	**71.42**	**1.29**	**64.32**	**1.42**

All experiments are running on a workstation with Intel Quad-Core 3.07 GHz CPU and 24 GB memory. The algorithms are implemented on MATLAB platform. Algorithm 1 usually converges in 100 iterations. It takes about 20 seconds to process one image given superpixel features extracted offline. We show several exemplar comparisons of segmentation results on BSD [15] in Figures 3, where each row shows the original image in column 1 and corresponding segmentation results obtained by MsLRR-IV, MeanShift [7] and MNCut [6] in columns 2, 3 and 4, respectively. Different semantic regions are indicated with different colors.

4.3. Exp-II: Internal Replication Prior and External Image Statistics

We further evaluate the effectiveness of the discovered replication prior by comparing it with the external image statistics. As aforementioned, there are several external statistics based methods, and we choose to implement the most recent one by Liu *et al.* in [13]. They proposed to extract superpixel co-occurrence frequencies from an extra unlabeled image corpus and utilize this knowledge for image segmentation. Our formulation in (11) can integrate this knowledge by re-defining the matrix Q^s as $Q^s_{ij} = \exp\{-c_{ij}\}$ where c_{ij} indicates the co-occurrence frequency of superpixels i and j. The details of calculating superpixel co-occurrence from unlabeled images are referred to the paper [13]. We call this implementation as MsLRR-V, which has the same setting as MsLRR-IV except that the definitions of Q^s are different.

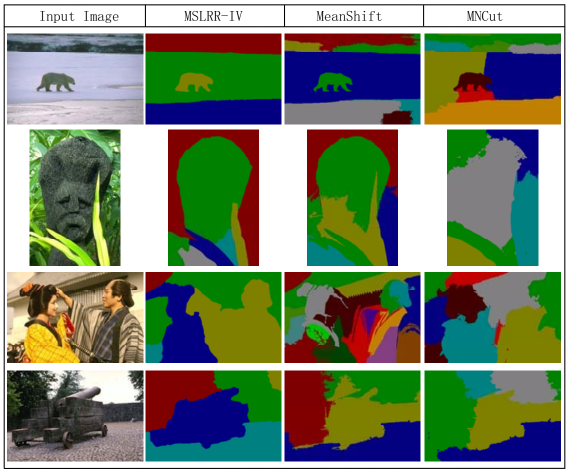

Figure 3. Exemplar comparisons of results by various methods on BSD500 [15]

We evaluate the algorithms MsLRR-IV and MsLRR-V on the test subsets of MSRC [19] and BSD500 [15] (see the introductions in Exp-I). The training subsets are used as the unlabeled image corpus for extracting the superpixel co-occurrence, as in [13]. In addition, we also use the unlabeled image corpus collected from Internet by Liu *et al.* [13], which contains 2300 images. All images are over-segmented into superpixels [16], and we use the same superpixel descriptors as in Exp-I.

Table 1 (b) and (c) report the quantitative results by MsLRR-V. From the comparisons in Table 1, one can observe that MsLRR-IV is able to achieve comparable accuracies as MsLRR-V that uses extra unlabeled images. To achieve the equally good quality of segmentation, MsLRR-V requires hundreds of unlabeled images. Furthermore, in BSD500 database, MsLRR-V using 300 unlabeled images (in Table 1-b) is inferior to the same algorithm using 2300 unlabeled images (in Table 1-c). We can obtain similar observations in MSRC dataset. These comparisons partially demonstrate that, external image statistics requires fairly large-scale unlabeled image corpus to achieve the robustness, which is consistent with the claims in previous works [13] [23]. In contrast, the proposed replication prior is extracted from the image itself, and thus has less limitations in real applications.

5. Summary

In this work, we studied a simple yet efficient internal image statistics and presented a practical method, the multiscale low-rank representation (MsLRR), for image segmentation. MsLRR aims to infer the low-rank refined superpixel affinity matrices at different scales of the input image in parallel, and meanwhile, to impose the cross-scale constraint to make the desired affinity matrices consistent. Extensive experiments on image databases well demonstrate the effectiveness of the proposed method.

6. Acknowledgement

This work was supported by the Army Research Laboratory 62250-CS, the National Institute of Heath 79592, and the Korean Ministry of Education, Science, and Technology (under the National Research Foundation WCU program R31-10008). Liang Lin was supported by the National Natural Science Foundation of China (Grant no.61173082), the Guangdong Science and Technology Program (Grant no. 2012B031500006), and the Guangdong Natural Science Foundation (Grant no. S2011010001378).

References

[1] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *CVPR*, 2009. 6

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, Contour Detection and Hierarchical Image Segmentation, *TPAMI*, 2011. 6, 7

[3] A. Buadess, B. Coll, and J. Morel. A non local algorithm for image denoising. In *CVPR*, 2005. 2

[4] S. Bagon and O. Boiman and M. Irani. What is a Good Image Segment? A unified Approach to Segment Extraction. In *ECCV*, 2008. 2

[5] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM.* , 2011. 2

[6] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *CVPR*, 2005. 2, 3, 4, 6, 7

[7] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *TPAMI*, 2002. 2, 6, 7

[8] A. Efros and T. Leung. Texture synthesis by nonparamaetric sampling. In *ICCV*, 1999. 2

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively trained Part Based Models. *TPAMI*, 2010. 3

[10] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV*, 2009. 2

[11] X.He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labelings. In *CVPR*, 2004. 6

[12] Z. Lin, A. Ganesh, J. Wright, M. Chen, L. Wu, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *SIAM Journal on Optimization*, 2011. 2, 5, 6

[13] X. Liu, J. Feng, S. Yan, L. Lin, and H. Jin. Segment an Image by Looking into an Image Corpus. In *CVPR*, 2011. 1, 2, 7

[14] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010. 1, 2, 3

[15] D. Martin, C. Fowlkes, D. Tal and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *ICCV*, 2001. 2, 3, 4, 6, 7

[16] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003. 1, 2, 6, 7

[17] B. Russell, A. Efros, J. William, F. Freemand and A. Zisserman. Segmenting scenes by matching images composites. In *NIPS*, 2009. 1, 2

[18] S. Rao, H. Mobahi, A. Y. Yang, S. Sastry and Y. Ma. Nat- ural image segmentation with adaptive texture and boundary encoding. In *ACCV*, 2009 6, 7

[19] 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. 2, 6, 7

[20] Z. Tu and S. Zhu. Image segmentation by data-driven markov chain monte carlo. *TPAMI*, 2002. 2

[21] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, 2009. 2

[22] A. Yang, J. Wright, Y. Ma and S. Sasty. Unsupervised segmentation of natural images via lossy data compression *CVIU*, 2008 6, 7

[23] M. Zonta and M. Irani. Internal statistics of a single natural image. In *CVPR*, 2011. 2, 7