MsLRR: A Unified Multi-scale Low-rank Representation for Image Segmentation

Xiaobai Liu\textsuperscript{1,2}, Qian Xu\textsuperscript{3}, Jiayi Ma\textsuperscript{1}, and Hai Jin\textsuperscript{1}
\textsuperscript{1}SCTS & CGCL, Huazhong University of Science and Technology, Wuhan, China
\textsuperscript{2}University of California, Los Angeles, CA
\textsuperscript{3}San Diego State University, San Diego, CA

Abstract—In this work, we present an efficient multi-scale low-rank representation (MsLRR) 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 noise, 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. The internal image statistics are referred to as replication prior, and we quantitatively justified 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. The proposed representation can be used for both unsupervised or supervised image segmentation tasks. Our experiments on public datasets demonstrate the presented method can substantially improve segmentation accuracy.

Index Terms—Low-rank Refined Affinity, Image Segmentation, Internal Image Statistics, Multi-scale Image Representation

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 past literature, many efforts have been devoted to extra information beyond the input image to solve this problem [20], [15]. 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 [20]. Second, fully exploring the cross-image statistics is obviously time-consuming and computationally inefficient, which also limits the application of these methods.

In this work, we introduce a simple yet efficient internal image statistic 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 [19], and construct an affinity graph by taking superpixels as graph vertices. Each superpixel is represented as one appropriate appearance feature, e.g., color histogram. 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 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 previous work [16], the low-rank constraint is helpful to suppressing the effects of data noise and corruptions.

The inference of low-rank refined affinity matrices serves two more purposes. The first one is a discriminative observation on natural images: small-size image patches (e.g. $6 \times 6$ pixels) with certain appearance patterns tend to recur frequently within the same semantic region, but may not appear in semantically different regions. We refer
this internal statistic to 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 are semantically similar. Generally, if every patch within one subregion has many copies in another one, these two subregions have high replication prior and they will belong to the same semantic region with high probability, and vice versa. In later sections, we will further quantitatively justify the above observation 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 [19].

The second purpose lies in 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 nuclear norm and $\ell_1$-norm minimization problem. This formulation is a convex quadratic function and we introduce an efficient optimization procedure based on the augmented Lagrange multiplier (ALM) method [14]. Taking the solved low-ranked refined superpixel affinities, one can call the Normalized Cut method [6] or the CRF method [22] to address the unsupervised or the supervised segmentation problems.

The contributions of this work are two-fold. First, we develop a multi-scale low-rank representation to seek the affinity matrix at each scale in parallel, while preserving the cross-scale consistency. Second, we study a simple yet efficient procedure to compute internal image statistics, and present a practical method to utilize this knowledge for real images. The advantages of our approach are demonstrated by extensive experiments with comparisons to the state-of-the-art segmentation algorithms on two public datasets MSRC [22] and BSD-500 [18].

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 (Markov Chain Monte Carlo) algorithm [23], Comaniciu and Meer’s Mean-shift [7], and Shi and Malik’s Normalized Cuts method (NCut) and its extension multi-scale NCut [6]. All of these three algorithms directly utilize the low-level feature descriptors which make the segmentation task painfully unconstrained. Alternatively, our method aims to solve for the low-rank refined affinity matrix while utilizing both the low-level features and the discriminative prior.

As aforementioned, extra knowledge or statistics have been studied to address the ill-posed nature of image segmentation [15] [20]. Nevertheless, these algorithms usually require 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 statistic, which has the obvious benefit of low computational demand. Moreover, we will experimentally show that, to achieve equally good quality of segmentation by the low-rank refined replication prior, hundreds of images are required for the external statistics [15].

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 [29] 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 statistics for image segmentation.

Our method is also motivated by the advances in subspace clustering, especially the low-rank representation (LRR) method, which seeks 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 [16]. The advantages of LRR have been well demonstrated in comparisons to other subspace clustering methods, e.g. sparse subspace clustering [27], especially while encountering severe corruptions and noise. Our method extends the LRR method to infer multiple low-rank representations separately at multiple scales, while simultaneously preserving 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 [19]. Each superpixel comprises an ensemble of pixels that are spatially coherent and perceptually similar with respect to certain appearance features (e.g. color). Let $m$ denote the number of scales, $I^s$ be the scaled image indexed by $s = 1, \ldots, m$. 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 two associated superpixels belonging to the same semantic region. Thus, the overall quality of segmentation depends on the pairwise superpixel affinity matrix. Let $x_i^s \in \mathcal{R}^d$ denote the $d$-dimension feature descriptor extracted for the $i$th superpixel in $I^s$. $\mathcal{R}$ 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 noise and clutter, the obtained affinity matrix is always corrupted. 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 produce 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 [16]. We aim to represent each superpixel descriptor as a linear combination of other superpixel descriptors, and seek the lowest rank representation of all superpixels in a joint fashion. Let $n^s$ denote the number of superpixels in $I^s$, and $X^s = [x_1^s, x_2^s, \ldots, x_n^s] \in \mathcal{R}^{d \times n^s}$. Each vector $x_i^s$ can be represented as the linear combination of the column vectors of $X^s$, denoted as,

$$x_i^s = X^s z_i^s,$$

(1)

where $z_i^s \in \mathcal{R}^{n^s}$ is the coefficient vector and $z_{ij}^s$ is the $j$-th component. 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 $Z^s = [z_1^s, z_2^s, \ldots, z_{n^s}^s]$ 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 $X^s$ to be corrupted. Padding the coefficient vectors into matrix form, we have,

$$X^s = X^s Z^s + E^s,$$

(2)

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

$$\min_{\{Z^s\},\{E^s\}} \sum_{s=1}^{m} \|Z^s\|_* + \lambda \|E^s\|_1$$

s.t. \hspace{1em} \text{s.t.} \hspace{1em} \|X^s - X^s Z^s + E^s\|_1$$

(3)

where $\| \cdot \|_*$ denotes the nuclear norm of a matrix, $\| \cdot \|_1$ denotes the $\ell_1$-norm of a matrix, and the parameter $\lambda > 0$ is used to balance the effects of the two parts. The $\ell_1$-norm in Eq. (3) is used to enforce most of the elements of $E^s$ have zero errors, i.e. only part of the data vectors are corrupted and others are clean. This is intuitive reasonable, especially when we choose to use the state-of-the-art superpixel descriptors that have high discriminative capability.

2.2 Objective-II: Replication Prior

The discovered replication prior is from a statistical observation on natural images: local small-size patches (e.g. $6 \times 6$ 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 that one superpixel will 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 $y_p$. Let $\Lambda$ denote a subregion and $q$ be the patch index in $\Lambda$. The empirical density of $p$ with respect to $\Lambda$ can be estimated using the Parzen window method:

$$D(p, \Lambda) = \frac{1}{|\Lambda|} \sum_{q \in \Lambda} \delta(K(||y_p - y_q||^2), \zeta)$$

(4)

where $K$ is a Gaussian kernel, and $|\Lambda|$ denotes the number of patches in $\Lambda$. $\delta(\cdot, \zeta)$ returns the first parameter if it is larger than $\zeta$, 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 $\zeta$ (fixed to be 0.4 in this paper) to depress the effects of partially similar patches.

We perform an experiment on the Berkeley Segmentation Database [18] to justify the replication prior. We use the test subset of 200 images. Only the original scale (about $480 \times 320$ 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 $\Lambda$ 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 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. Fig. 2 plots the density comparisons while looking at different patch-sizes, including $6 \times 6$, $10 \times 10$, $12 \times 12$ and $14 \times 14$ 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 evident while using fairly small-size patches.

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

$$Q^s_{ij} = e^{-(\frac{1}{\gamma} \sum_{p\in \Lambda^s_i} D(p,\Lambda^s_j) + \frac{1}{2\gamma} \sum_{q\in \Lambda^s_j} D(q,\Lambda^s_i))}.$$  

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

$$\min_{\{Z^s\},\{E^s\}} \sum_s \|Z^s\|^2 + \lambda \|E^s\|_1 + \beta tr(Z^sTQ^s)$$

s.t. $X^s = X^sZ + E^s$  

where $Z^sT$ indicates the transpose of $Z^s$, $tr(\cdot)$ returns the matrix trace, and $\beta$ is a tuning parameter. Minimizing the term $tr(Z^sTQ^s)$ will encourage the intra-class superpixel-pairs to have higher affinities (in $Z^s$) than the inter-class superpixel-pairs. In this way, the 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 high-level semantic statistics, which, although estimated using low-level appearance features, are able to describe how the small-size patches form the large-size semantic regions and thus provide 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-scale to the fine-scale 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 denote $I^{s+1}_j$ 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-scale, their affinity should be equal to the average affinities between their respective children at the fine-scale. Fig. 3 illustrates this constraints.

Moreover, we arrange all elements of $Z^s$ in one column vector

$$f(Z^s) = [Z^s_{i1} \ Z^s_{i2} \ ... \ Z^s_{in^s}]^T \in \mathbb{R}^{(n^s\times n^s)\times 1}.$$  

Fig. 2. Discriminative Replication Prior. We compare the mean intra-class patch densities and mean inter-class patch densities under different patch sizes, including $6 \times 6$, $10 \times 10$, $12 \times 12$, and $14 \times 14$ pixels. The images are from BSD database [18]. See texts for more details.
Let \( h^{s+1}(I_a) \) indicate the index of the superpixel in \( I^{s+1} \) which is the parent of \( I_a \). Let \( k = < a, b > \) index the elements or superpixel-pairs of \( f(Z^s) \) that are formed by \( I_a \) and \( I_b \), and \( l = c, d > l \) 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.

1. Set \( C^{s+1,s} \in R^{(n^{s+1} \times n^{s+1}) \times (n^{s} \times n^{s})} = 0 \);
2. For \( \forall k = < a, b >, k = 1, ..., n^s \); \( \forall l = c, d >, l = 1, ..., n^{s+1} \),
   - if \( h^{s+1}(I_a) = h^{s+1}(I_b) \),
     - set \( C^{s+1,s}(l, k) = 0 \);
   - else
     - if \( (h^{s+1}(I_a) = c, h^{s+1}(I_b) = d) \) or \( (h^{s+1}(I_a) = d, h^{s+1}(I_b) = c) \),
       - set \( C^{s+1,s}(l, k) = 1 \);
3. 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 squares function,

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

Furthermore, let \( Z = [f(Z^1); f(Z^2), \ldots, f(Z^m)] \), and

\[
C = \begin{bmatrix}
C_{1,1} & -1^2 & 0 \\
0 & \ddots & \ddots \\
0 & \cdots & C_{m,m-1} & -1^m
\end{bmatrix}. \tag{9}
\]

where \( 1^s \) 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 that the desired superpixel affinity matrices at different scales are consistent with each other.

\[ Z^2_{1,2} = \frac{1}{4} (Z^1_{1,3} + Z^1_{1,4} + Z^1_{2,3} + Z^1_{2,4}) \]

**Fig. 3.** Illustration of the cross-scale consistency constraint. The superpixel \( I^2_2 \) at the scale \( s=2 \) is the parent of the superpixels \( I^1_1 \) and \( I^1_3 \), and \( I^2_2 \) is the parent of \( I^1_1 \) and \( I^1_4 \). This leads to one constraint shown in the bottom line. See texts for more details.

**Algorithm 1 Optimization Procedure to Eq. (12)**

**Input:** data matrix \( \{X^s\}, s = 1, ..., m \), parameter \( \lambda, \beta \) and \( \gamma \);

Set \( J^s = Z^s = P^s = E^s = 0, Y^s = 0, V^s = 0, U^s = 0, \mu = 10^{-6}; \varepsilon = 10^{-5}, \rho = 1.1 \);

**While not converged do (iteration body),**

1. Update \( J^s \) by solving

   \[
   J^s = \arg \min_{J^s} \frac{1}{\mu} \| J^s \|_1 + \frac{1}{2} \| J^s - (Z^s + V^s) / \mu \|^2
   \]

2. Update \( P^s \) by \( P^s = (X^sT X^s + 1)^{-1}[X^s T X^s - X^s T E^s + Z^s + (X^s T Y^s - \beta Q^s + U^s) / \mu] \);

3. Update \( Z \) by \( Z = (\varepsilon C^T C + 1)^{-1} \lambda B / 2 \), where the vector \( B \) has following structure:

   \[
   B = [ f(G^1)^T, f(G^2)^T, \ldots, f(G^m)^T ]^T
   \]

   with \( G^s = J^s + P^s - (V^s + U^s) / \mu \), \( s = 1, \ldots, m \);

4. Update \( \{E^s\} \) by solving

   \[
   E^s = \arg \min_{E^s} \frac{1}{\mu} \| E^s \|_1 + \frac{1}{2} \| E^s - (X^s - X^s P^s + Y^s) / \mu \|^2
   \]

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: \( \forall s, \| X^s - X^s Z^s - E^s \| < \varepsilon \);

**2.4 Unified Formulation**

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

\[
\min_{\{Z^s\}, \{E^s\}} \sum_{s=1}^{m} \| Z^s \|_1 + \lambda \| E^s \|_1 + \beta \text{tr}(Z^s T Q^s) + \gamma \| CZ \|^2
\]

s.t. \( X^s = X^s Z^s + E^s \)

where \( \gamma \) is a tuning parameter. We will introduce the choices of these three parameters \( \lambda, \beta \) and \( \gamma \) in the section.
of experiments.

2.5 Optimization
To optimize Eq. (11), we adopt the Augmented Lagrange Multiplier (ALM) method [14] due to its efficiency in convergence. By introducing the auxiliary variables \( \{J^s\}, \{P^s\}, s = 1, \ldots, m \), we convert Eq. (11) to an equivalent problem,

\[
\min \sum_{s=1}^{m} \|J^s\|_* + \lambda \|E^s\|_1 + \beta \text{tr}(P^sTQ^s) + \gamma \|CZ\|^2
\]

s.t. \( X^s = X^sP^s + E^s, Z^s = J^s, Z^s = P^s \) \( (12) \)

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

\[
\min \gamma \|CZ\|^2 + \sum_{s} \|J^s\|_* + \lambda \|E^s\|_1 + \beta \text{tr}(P^sTQ^s)
+ \left< Y^s, X^s - X^sP^s - E^s \right> + \mu \|X^s - X^sP^s - E^s\|^2
+ \left< V^s, Z^s - J^s \right> + \frac{\mu}{2} \|Z^s - J^s\|^2
+ \left< U^s, Z^s - P^s \right> + \frac{\mu}{2} \|Z^s - P^s\|^2,
\]

(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) with respect to \( \{Z^s\}, \{E^s\}, \{J^s\} \), 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 [14], while \( \mu \) 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 [14]; ii) the subproblems in Step-3 can be solved analytically; and iii) the update factor \( \rho \) is set to be 1.1 to obtain a series of increasing \( \mu \).

3 IMPLEMENTATION
We apply the proposed low-rank refined image representation for image segmentation under both unsupervised and supervised settings.

3.1 Unsupervised Image Segmentation
Suppose the optimal superpixel affinity matrices are solved from Eq. (11), denoted as \( Z^s, s = 1, \ldots, m \). 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:

\[
W^s_{ij} = \sum_{k=1}^{3} w^k K^k_{ij}
\]

where \( K^k \) is a Gaussian kernel in a specific feature space, \( w^k \) is linear combination weight. Here, \( i,j \) index the pixels in images. In particular, \( K^3 \) is an appearance kernel, defined as \( K^3_{ij} = \exp(-\alpha^1 \Delta^a_{ij} - \alpha^2 \Delta^s_{ij}) \). \( \Delta^a_{ij} \) indicates the Euclidean distance between the color vectors at pixel \( i \) and pixel \( j \). \( \Delta^s_{ij} \) indicates the geometrical distance between pixel \( i \) and \( j \). \( K^2 \) is the smoothness kernel, defined as \( K^2_{ij} = \exp(-\alpha^3 \Delta^a_{ij}) \). These two kernels, following previous works [6], [22], [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:

\[ K^3_{ij} = \exp(-\alpha^4 \Delta^a_{ij}), \text{ where } \Delta^a_{ij} = \exp(-s^a_{ij}/\sigma^2) \]

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 \( W^s_{ij}, s = 1,2, \ldots \), to achieve the final segmentation.

3.2 Supervised Image Segmentation
We adopted the CRF framework [22] [12] to address the supervised image segmentation task. We first apply Algorithm 1 over the test image to extract its low-rank refined superpixel affinities, and further combine it with the CRF model trained from a set of training samples. Let \( \mathcal{X} \) denote the random field over a set of variables \( v = \{v_1, v_2, \ldots\} \) where each variable corresponds to one pixel in images. The domain of each variable is a set of image labels \( \mathcal{L} = \{l_1, l_2, \ldots\} \). The corresponding Gibbs energy is composed of a set of unary and pairwise potentials,

\[
E(v) = \sum_i \phi^u(v_i) + \sum_{i,j} \phi^p(v_i, v_j),
\]

(15)

where \( i \) and \( j \) range from 1 to the number of variables. The unary potential \( \phi^u \) is calculated separately for each pixel by a classifier that produces a distribution over the label assignment given image features. We use the implementation in [22] that incorporates shape, texture, location and color descriptors to train classifiers, as described in Section 4. The pairwise potentials in our method take the form:

\[
\phi^p(v_i, v_j) = \varphi(v_i, v_j) \sum_{k=1}^{3} w^k K^k_{ij},
\]

(16)

where \( \mu \) is a label compatibility function. We use \( \varphi(v_i, v_j) = |v_i \neq v_j| \) to introduce a simply penalty for nearby similar pixels that are assigned different labels. \( w^k \) and \( K^k \) are defined in Eq (14) to combine the smoothness in terms of geometric, appearance, and replication prior. Parameters are learnt from training data as described in Section 4.

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 [22] and Berkeley Segmentation Database (BSD) [18]. 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 [19]. 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 [22]. These descriptors are first normalized by their respective $\ell_2$-norms and then concatenated to form one single descriptor vector (i.e. $\chi^i$). Once obtained the feature matrix $X^i$, we perform PCA [25] on it, and remove the zero principle components to obtain a low-dimensional feature matrix. Same feature descriptors are extracted while computing matrix $Q^i$ in Eq. 5.

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 VI values indicate better performance.

4.2 Exp-I: Unsupervised Image Segmentation

We first 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 the low-rank refined affinity matrices at multiple scales separately. 2) MsLRR-II, that sets $\beta = 0$, and only uses the regularization term of $\gamma \|Z\|_F^2$. 3) MsLRR-III, that sets $\gamma = 0$, and only uses the regularization term of $\beta tr(Z^T Q^s)$. 4) MsLRR-IV, that utilizes both regularization terms. We also implement two popular image segmentation 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 $\Delta_{ij}^s = \exp(-|x_i - x_j|^2/2\sigma^2)$ in Eq. (14) (see Subsection 3.1), and call MNCut to segment images. We refer this algorithm to as MNCut-SP and the original method [6] to as MNCut. We perform the various algorithms under a variety of scale parameters. This evaluation strategy is also used in previous works, e.g., [2] [21]. For MsLRR and MNCut based algorithms, the number of segments is set within $\{2, 3, \ldots, 15\}$. The minimum region area of MeanShift algorithm is set within $\{1500, 2000, \ldots, 10000\}$ pixels. For MsLRR algorithms, we set $\beta = 0.01$, and $\gamma = 0.01$ if the related regularization term is used. Set $\lambda = 0.01$. The patch-size is fixed to be $6 \times 6$ pixels. The threshold $\zeta$ 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 $\sigma = 1$. Although we set the above parameters empirically, in practice, our method can produce similar segmentation results.

Table 1(a) reports the performance comparisons of the various unsupervised segmentation algorithms on the MSRC [22] and the BSD [18] databases. We also include the available results of other state-of-the-art segmentation algorithms for comparisons, that refer to as UCM [2], TBES [21] and CTM [28]. We can make the following observations. 1) The proposed MsLRR algorithms can achieve comparable accuracies with the state-of-the-art methods UCM [2], and TBES [21], and much better results than other baseline methods [2]. 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.

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 are extracted offline. Several exemplar comparisons of segmentation results are illustrated in Fig. 4, 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. These comparisons under various conditions well validate the effectiveness of our approach.

We further evaluate the effectiveness of the discovered
replication prior by comparing it with the external image statistics. As aforementioned, there are several external statistical based methods, and we use the method in our previous work [15], which extracted superpixel co-occurrence frequencies from an extra unlabeled image corpus (namely images are given without any annotations) and utilized this knowledge for image segmentation. Our formulation in Eq. (11) can integrate this knowledge by re-defining the matrix $Q_{ij} = \exp(-c_{ij})$ 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 [15]. We denote this implementation as MsLRR-V, which has the same setting as MsLRR-IV except that the definitions of $Q$ are different.

We evaluate the algorithms MsLRR-IV and MsLRR-V on the test subsets of MSRC [22] and BSD500 [18]. The training subsets are used as the unlabeled image corpus for extracting the superpixel co-occurrence, as in [15]. In addition, we also use the unlabeled image corpus collected from Internet in [15], which contains 2300 images. All images are over-segmented into superpixels [19].

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 with MsLRR-V that uses extra unlabeled images. To achieve equally good quality of segmentation, MsLRR-V requires hundreds of unlabeled images. Furthermore, in the 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 make similar observations in MSRC dataset. These comparisons demonstrate that, external image statistics require a fairly large-scale unlabeled image corpus to achieve the robustness, which is consistent with the claims in previous works [15] [29]. In contrast, the proposed replication prior is extracted from the image itself, and thus has less limitations in real applications.

4.3 Exp-II: Supervised Image Segmentation

We further evaluate the presented image representation under the supervised setting. The unary potentials are derived from TextonBoost [22], that contains 17-D filter bank, color, histogram of oriented gradients, and pixel location features. We learn the parameters $w^1, w^2, \alpha^1, \alpha^2, \alpha^3, \sigma$ by grid search on a validation set (built from the training images) [22]. The smoothness parameters $w^2, \alpha^3$ are set to be 0.3, 0.1 that work well in practice. The inference algorithm is the same as the work [22]. We compare our algorithm to six supervised segmentation algorithm on MSRC [22], refer them to as TextonBoost [22], HIM [30], Auto-Context [24], PLSA-MRF [26], H-CRF[12] and Dense-CRF [13]. Table 2 reports the comparisons in term of Covering Rate (CR) between the above algorithms, where the figures of baseline methods are directly from their respect literatures. Our method clearly outperforms all baselines except the Dense-CRF [13], which additionally utilizes a logistic regression classifier on the responses of the boosted classifiers for training unary terms. These comparisons demonstrate that our method can achieve the state-of-the-art segmentation accuracies.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextonBoost [22]</td>
<td>72.9</td>
</tr>
<tr>
<td>PLSA-MRF [26]</td>
<td>73.5</td>
</tr>
<tr>
<td>Auto-Context [24]</td>
<td>77.7</td>
</tr>
<tr>
<td>HIM [30]</td>
<td>81.2</td>
</tr>
<tr>
<td>H-CRF[12]</td>
<td>84.5</td>
</tr>
<tr>
<td>Dense-CRF [13]</td>
<td>86.0</td>
</tr>
<tr>
<td>MsLRR</td>
<td>85.3</td>
</tr>
</tbody>
</table>

5 Summary

In this work, we studied a simple yet efficient internal image statistic and presented a practical method, the multi-scale low-rank representation (MsLRR), for image segmentation. MsLRR infers low-rank refined superpixel affinity matrices at different scales of the input image in parallel, and imposes the cross-scale constraint to make the desired affinity matrices consistent. Extensive experiments conducted with image databases demonstrate the effectiveness of the proposed method. Although we focus on the region grouping problem in this work, the developed formulations and prior model can be easily extended for other high-level image tasks, e.g. semantic contour detection, which will be further exploited in future.

References


E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? J. ACM., 2011.


P. Krahenbuhl and V. Koltun Efficient Inference in Fully connected CRFs with Gaussian Edge Potentials. In NIPS, 2011


Xiaobai Liu is currently a Research Scholar in the Department of Statistics, University of California at Los Angeles (UCLA). He received his Ph.D from the school of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China in 2011. Before that, he had been working in the Dept. Electric & Computer Engineering, National University of Singapore and working as the Research Associate for Professor Shuicheng Yan. He also spent years as Research Associate at Lotus Hill Institute under the supervision of Professor Song-Chun Zhu from 2007 to 2008. He has published more than 20 articles over a series of research topics and now his research interests fall on computer vision, machine learning and large scale image retrieval etc.

Qian Xu received the B.S. degree from the School of Science, Beihang University (BUAA), Beijing, China, in 2006 and M.S. degree from the Department of Mathematics and Statistics, San Diego State University (SDSU), CA, in 2011. She is currently a Ph.D. student of the Department of Mathematics and Statistics, SDSU. Her research interest fall in the typical nonparametric models and their applications in image data.
Jiayi Ma received the B.S. degree from the Department of Mathematics, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2008. He is currently a Ph.D. student of the School of Automation, HUST. From September 2012 to September 2013, he was with the Department of Statistics, UCLA, Los Angeles, CA. His research interests are in the areas of computer vision and machine learning.

Hai Jin is a professor of computer science and engineering at the Huazhong University of Science and Technology (HUST) in China. He is now Dean of the School of Computer Science and Technology at HUST. Jin received his PhD in computer engineering from HUST in 1994. In 1996, he was awarded a German Academic Exchange Service fellowship to visit the Technical University of Chemnitz in Germany. Jin worked at The University of Hong Kong between 1998 and 2000, and as a visiting scholar at the University of Southern California between 1999 and 2000. He was awarded Excellent Youth Award from the National Science Foundation of China in 2001. Jin is the chief scientist of ChinaGrid, the largest grid computing project in China.

Jin is a senior member of the IEEE and a member of the ACM. Jin is the member of Grid Forum Steering Group (GFSG). He has co-authored 15 books and published over 400 research papers. His research interests include computer architecture, virtualization technology, cluster computing and grid computing, peer-to-peer computing, network storage, and network security.
Jin is the steering committee chair of International Conference on Grid and Pervasive Computing (GPC), Asia-Pacific Services Computing Conference (APSCC). Jin is a member of the steering committee of the IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid), the IFIP International Conference on Network and Parallel Computing (NPC), and the International Conference on Grid and Cooperative Computing (GCC), International Conference on Autonomic and Trusted Computing (ATC), International Conference on Ubiquitous Intelligence and Computing (UIC)