Highlights

• A new unsupervised shape discovery method using a novel joint foreground/background segmentation and a dense part-part correspondence between all image pairs.

• A new multiscale spectral synchronization method which jointly align the spectral representations of all input images.

• A novel unsupervised superpixel-based groupwise co-registration which converts the implicit eigenvector-eigenvector synchronization to superpixel-superpixel dense correspondences.
Unsupervised Shape Discovery using Synchronized Spectral Networks

Yunliang Cai, Andrea Lum, Ashley Mercado, Mark Landis, James Warrington, Shuo Li∗

the Digital Imaging Group of London
Dept. of Medical Imaging, Western University
268 Grosvenor St., SJHC, London, Ontario, N6A 4V2, Canada

Abstract

Unsupervised discovery and extraction of common shapes from unlabeled images is a fundamental problem in object recognition and has broad applications in practice. However, shape discovery suffers from the lack of consistent matching methods for finding the correspondences between objects with different colors/textures among the input images. In this paper, we propose a novel unsupervised shape discovery method using Synchronized Spectral Network (SSN) which provides automatic part-part correspondences across images. The SSN is a spectral graph-based model that encodes the pixel self-similarities of different images in spectral bases, and synchronizes the bases between images to achieve the part-part correspondences. Unlike explicit feature matching, correspondences obtained by spectral synchronization are independent of colors/textures and image modalities. An image network can then be built by spectral correspondences where the common shapes among them can be easily identified and segmented. Our results in multiple shape discovery datasets demonstrate that we outperform the state-of-the-art object/shape discovery methods, providing better segmentations for common shapes.

Keywords: Spectral synchronization, joint image matching, groupwise segmentation, shape discovery

∗Corresponding author
Email address: slishuo@gmail.com (Shuo Li)
1. Introduction

Shape is an important cue for object recognition which is invariant for image modality, shading, color, and texture appearance. Shape matching lays the foundation of numerous successful computer vision algorithms from image segmentation to high-level object categorization. The understanding of shape representation, shape correspondence, and the shape formation can help tackle the fundamental challenge of object recognition. Among all the shape problems in computer vision, unsupervised discovery of object shapes is one of the most challenging. Solving the unsupervised shape discovery can lead to broad applications in practice and enable the complete automation of object segmentation, classification, detection, and categorization. However, the common shape among images is difficult to extract not only because the prior information about the object is simply unknown, but also because of the contamination of object shapes with colors, textures, and illuminations in various image backgrounds.

Shape discovery is closely related to the study of object discovery [1] but differs from their objectives and methodologies. Unlike the object discovery which aims on finding the common object pattern categories and their locations (i.e., [2]), shape discovery aim to extract the shape outlines (i.e., [3]). This makes shape discovery more compatible in some image representations, i.e., sketch images in forensic study. Because of their different objectives, existing object discovery methods search the matching of image blocks and the block-combined patterns while shape discovery search for the object boundary and/or the continuously combined edge fragments for the shape.

Existing unsupervised discovery is often considered as a clustering problem for a number of shape-related features such as boundaries and local patches [4] [3] [5] [6] [7]. The recovered shapes are represented as contour fragments in these methods. However, contours are not sufficient to extract the object shapes out of the complex image backgrounds unless strong supervised models are applied. Also, these methods lack the consistent part-part correspondences.
for the common object shapes across images. The final segmented shapes are often contaminated with unrelated image parts.

**Shape Discovery via Consistent Part-Part Correspondences.** To overcome the inconsistent matching problem in shape discovery, we aim to build up part-part correspondences across images. Comparing to most current shape discovery methods, which mainly focus on contour or texture matching, part-part correspondences can transfer and re-enforce the shape patterns from one image to another. By accumulating the correspondence cues in the images, the repetition of common shapes can be easily identified and extracted. In addition, the resulting shape segmentations under part-part correspondences naturally provide sub-part correspondences among the shapes. This leads to the discovery of not only the common shapes in the images but also their intrinsic hierarchical components, which will be more useful in object recognition than simple object identification/localization in object discovery problem (i.e., [8]). Particularly, the multiscale subparts correspondences provide joint labeling of object parts for input images, which improves the accuracy of object classification and reveals more geometric information such as object poses, human behaviors, and facial expressions.

**Implicit Correspondence Matching.** Building explicit pixel-pixel correspondences among images are exhaustive and unstable. Instead of the explicit matching, we learn the part-part correspondences in the input images by clustering their intra-image self-similarity features. Image regions with same intra features are assigned to the same cluster, which leads to part-part correspondences for different images without explicit conducting matching like template convolution or correlation analysis. We use the extended Synchronized Spectral Networks (SSN) model [9] for extracting the intra image features. The model generates a set of quasi-eigenvectors and forces them to have similar appearances. The joint feature learning implicitly associates the similar image parts to the same feature cluster, and the shared common shapes can be discovered by analyzing the combinations of feature categories in image regions.

**Main Contribution.** We propose a new general unsupervised shape dis-
covery method using a novel SSN model.

- The proposed method extracts common shapes from a collection of unlabeled images with joint foreground/background segmentation and dense part-part correspondence between all image pairs. No manual annotation is needed for the whole process.
- The multiscale spectral synchronization of the SSN model is first applied to solve the groupwise matching of heterogeneous image collection. The SSN model converts the explicit pixel-pixel shape matching to an implicit eigenvector-eigenvector synchronization so that shape matching across different image modalities (color/greylevel/sketch) becomes possible.
- The new synchronized superpixel-based common shape sketch algorithm enables the grouping of synchronized region into shared common shapes, which allows the unsupervised discovery of complicated shape patterns such as horse, face, and airplane.
- The new superpixel transferring can propagate the common shapes shared among the small image sets to large image set. As a byproduct of the algorithm, the transferring generates an efficient SSN-based detector for searching the recurring objects.

2. Related Work

Object Discovery. The classical shape discovery originates from the object discovery problem [8] [10] [11] [12] [4], which aims to recognize, categorize, and locate objects or natural scenes in a set of unlabeled images. Object discovery methods often focus on the statistics or spatial distributions of the image features. The unsupervised learning of object often exploits probabilistic models such as Probabilistic Latent Semantic Analysis (PLSA) [10], Latent Dirichlet Allocation (LDA) [11], and spectral clustering [4] for grouping local image features into meaningful classes. The discovery methods can also reveal the hierarchical relations among the object class [13] [14], providing hierarchical categorization.
for the images. Although the object discovery can reveal the unknown object
classes/locations in unlabeled images, the discovery cannot provide detailed
subpart correspondences and thus cannot extract the object shapes.

**Unsupervised Shape Discovery.** One of the earliest unsupervised shape
discovery method was proposed by Lee and Grauman [3]. The authors proposed
a foreground shape discovery method by matching and clustering the contour
segments and SIFT features. Following the same idea, existing shape discov-
ery related methods focus on low-level vision features such as edges/contours
to identify the shape patterns in images. Payet and Todorovic [6] proposed a
shape-based object discovery method that relied on the pairwise matching of
contours using a special graph-based multi-label assignment model called Co-
proposed a bottom-up contour grouping approach for shape matching using the
many-to-one matching, a contour embedding method for extracting the salient
contours shared across images. The salient long contour matching became the
key ingredient for building the common shape sketch as presented in Marvaniya
et al [16]. The author used hierarchical matching to combine the long contours
shared across images into common shapes. The bottom-up shape clustering
approach can effectively identify the salient shape parts of the common object,
but the resulting shapes were often incomplete and spoiled with unrelated noisy
edges.

Instead of the bottom-up grouping of image features, Bagon et al [5] pro-
posed a global common shape extraction method using self-similarity descriptor
[17]. The sketch output of their method was a foreground mask representing
the common object shapes shared among the input images. However, the self-
similarity descriptor became vulnerable when the image set grew larger or the
shape patterns turned more complicated. The extracted mask will be blurred
due to the increasing variations in a large dataset. Rubinstein et al [18] [19] used
a warping-based dense correspondence matching ([20]) for generating a robust
neighbor set for each image. Images in the same neighbor set can be efficiently
co-segmented thanks to the recurring shape pattern between each other. De-
spite their limited performance in large data set, the global approach inspires us to use the correlation of global image structures other than the exhaustive bottom-up grouping of low-level vision features.

**Weakly Supervised Shape/Object Discovery.** Some recent studies of shape discovery suggested alleviating the hard unsupervised discovery by solving a weakly supervised problem instead. Unlike a supervised approach where explicit shape-labeled images were supposed to be used, bounding boxes or over-segmented regions will serve as the training samples to discover the common shape pattern in the weakly supervised approach. Lee and Grauman [21] used a set of segmented region obtained from feature-based classification to discover the shared structure in images. They applied the shared structure for the new top-down object segmentation with additional background and shape smoothness visual cues. Kang et al [22] proposed a similar shape-based segmentation grouping method specifically for the discovery of recurring objects for images from daily living scenes. The top-down shape discovery was also adopted in Chen et al [23] where the authors used unsupervised segmentation over a selected homogenous image set for generating the shape priors of arbitrary visual subcategories. The shape priors can then be applied to new object discovery and segmentation. On the basis of these successful weakly supervised methods, a possible way to achieve unsupervised shape discovery is to particularly automate the annotation stages in them, as the method proposed in this paper.

**Cosegmentation and Segmentation Propagation.** Another study that related to unsupervised shape discovery is the image cosegmentation [24] [25] [26] and segmentation propagation [18] [27]. The conventional cosegmentation method was more similar to the early unsupervised shape discovery methods where a bottom-up grouping for similar features across images was conducted to jointly produce the segmentations for all images. Cosegmentation was initially studied in Rother et al [24] for simultaneously segmenting the common parts of image pairs. The idea was further extended to multiple image segmentation. Joulin et al [26] combined the normalized cut (NCut) with the discriminative clustering (kernel SVM) for cosegmenting distinct image pairs. Kim et
al [25] formulated this cosegmentation problem as a submodular optimization and provided an efficient computation algorithm for the cosegmentation over large-scale image collections. The cosegmentation model inspired the study of segmentation propagation. Similar to the weakly supervised shape/object discovery methods, segmentation propagation trained segmented foreground masks in the seed images and propagate the masks to other images. An arbitrary image will be segmented under the guidance of its similar seed images. Rubinstein et al [18] used word tags and label maps to provide pixel-wise transfer through dense image correspondences. Guillaumin et al [28] exploited the hierarchical structure of ImageNet to transfer the window-level segmentation across images. Jain and Grauman [27] constructed a segmentation graph for the image set, and used the pairwise region similarity in the graph to select the most influential images. The selected images contain the shared shape patterns that can largely improve the object segmentation in the set.

**SSN Model.** The SSN model we employed in this paper was introduced in [9] for MR/CT cardiac segmentation. The proposed SSN-based shape discovery method has many new extensions for the previous MR/CT segmentation model: 1) new extension of the discovery and segmentation of general shapes with different scales and poses; 2) new extension of the multiple modality synchronization in real images environment (color, grey-level, sketch); 3) new extension of the grouping of synchronized superpixels for common shape extraction; 4) new extension of the iterative correspondence transfer for propagating the synchronized superpixels to large image set; 5) new extension of refined shape extraction using the guidance of common synchronized superpixels.

### 3. Method Overview

Suppose there is an image set where all images share a common shape but are distorted by different poses and mixed with various colors, textures, and image backgrounds. Our goal is to automatically extract the common shape in each image without any prior training processes.
Figure 1: The overview of our approach. For an input image set, first build up their superpixel correspondences using the SSN model. A common shape sketch is obtained by the clustering over the synchronized superpixels and then are refined by the detail GrabCut segmentation to extract the desired shapes. The extracted shapes will serve as detectors to iteratively discover and segment new shapes in other images. See Sec. 3 for general discussion of each step.

We consider an iterative approach to achieve the unsupervised discovery. We start the unsupervised learning in a small pool of images, obtaining the image correspondences and the common shapes. Using the initial unsupervised discovery, we then match the new image outside the pool and adaptively transfer the segmentation until all images are segmented. There are three steps in our method as shown in Figure 1:

Step 1. Building image part-part correspondences via SSN (Sec. 4). In this step, we search for part-part correspondences across different images. Unlike traditional matching schemes which use explicit image alignment/registration, we build up the correspondences by clustering the image regions that are sharing the same self-similarity features. The clustering is implemented by synchronizing the spectral graph bases over the images. We use a Synchronized Spectral Networks (SSN) to build up the part-part correspondence across images, and label the correspondences as Synchronized Superpixels (SSP).

Step 2. Shape discovery and segmentation from part-part correspondences (Sec. 5.1). With the image part-part correspondences, we extract the common foreground shape regions in images and obtain the final shape segmentations. The self-similarities among synchronized superpixels can be utilized for obtaining the shape correspondences among the images. The shape correspondence estimation is implemented by a common shape sketch algorithm [5], and the final shape segmentation is obtained by GrabCut [29].
Step 3. Iterative correspondence transfer (Sec. 5.2). For images with large shape scale differences and deformations, we can use the input images in Step 1 and 2 as ‘seed images’ to estimate and segment the new images. The new images are sent to spectral re-synchronization with the seed images in SSN. Through the re-synchronization, the part-part correspondences in seed images will transfer to the new images, and the new shapes can be identified and extracted using the same scheme in Step 2.

4. Synchronized Spectral Networks for Part-Part Image Correspondences

The SSN is a network formed by a set of spectral graphs that can share their structures with each other through a synchronization. We use the spectral bases to represent the structure of a graph. The spectral bases of all graphs in the network are synchronized to each other such that their structures are matched. Joint analysis of the images (i.e., segmentation, shape detection) becomes possible by simply exploiting the matched structures among them.

4.1. From Spectral Graph to Spectral Network

An image can encode its pixel self-similarities in a spectral graph. The set of spectral graphs can form a “graph of graphs”, which we call it as spectral network when these spectral graphs are correlated with each other.

For an image $I$, we build graph $G = (V, E)$ such that $V$ is the set of pixels and each edge $e \in E$ connects two pixels $i, j$ in the image. Suppose edge $e$ for pixel $i, j$ is weighted by $W(i, j)$, then $W$ is a $N \times N$ matrix for $N = |V|$. The weight between two pixels is determined by the pixel intensities and contour interventions:

$$W(i, j) = \exp(-||x_i - x_j||^2/\sigma_x - ||I_i - I_j||^2/\sigma_I - \max_{x \in \text{line}(i, j)} ||\text{Edge}(x)||^2/\sigma_E)$$ (1)

where $x_i, x_j$ are the location of the pixels $i, j$ and the $I_i, I_j$ are their intensities respectively. $\text{Edge}(x)$ represents an edge detector in location $x$. $\sigma_x, \sigma_I, \sigma_E$ are
Figure 2: Example of Synchronized Spectral Networks (SSN) for building part-part correspondences across images. SSN generates Synchronized Spectral Bases at each scale by forcing the eigenvectors of the image spectral graphs to match each other. By the K-means clustering of the synchronized spectral bases, the set of correlated parts from different images are labeled with the same color and assigned as a Synchronized Superpixel (SSP).

constants that will be assigned empirically. In practice, \( W(i,j) \) will only be computed in the set of \( k \)-nearest neighbors. Suppose image \( I \) contains \( N \) pixels, then \( W \) is a \( N \times N \) sparse matrix. Let \( D \) be the diagonal matrix whose elements are the row summations of \( W \). We can have the Laplacian matrix

\[
\mathcal{L} = \mathbf{I}_D - D^{-1/2} WD^{-1/2}
\]

where the eigenvectors of \( \mathcal{L} \) are the set of unsynchronized spectral bases. The Laplacian can thus be approximated by first \( K \) eigenvectors \( \mathcal{L} \approx \sum_{k=1}^{K} \lambda_k \xi_k \xi_k^T \) where \( \lambda_k \) and \( \xi_k \) are the eigenvalue and its associated eigenvector respectively. These spectral bases encode the self-similarity information described by (1).

4.2. Spectral Synchronization

Spectral synchronization is the process of forcing the spectral bases from one image to match the bases from the other image, so that for each spectral channel the spectral bases will have similar appearances. Once the spectral bases of two images are synchronized, the self-similarities features of both images are matched accordingly. This provides an implicit alignment for both images without explicitly computing the registration.

Joint Laplacian Diagonalization [30] [31] is a method to obtain pairwise matching among spectral bases. Suppose \( \mathcal{L}_m \) is the Laplacian matrixes for
I_m \in \mathcal{I}, the goal of the joint diagonalization is to obtain a new set of generalized quasi-eigenvectors \( Y_m = [y_{m,1} \ldots y_{m,K}] \in \mathbb{R}^{N \times K} \) which satisfies

\[
\sum_{m \in \mathcal{I}} ||Y_m^T L_m Y_m - \Lambda_m||_F^2 < \epsilon
\]

(3)

for \( \Lambda_m = \text{diag}(\lambda_1(\mathcal{G}_m), \ldots, \lambda_K(\mathcal{G}_m)) \) and \( \epsilon > 0 \) is small. In addition, different quasi-eigenvectors \( Y_m \) and \( Y_l \) should be matched in a feature space, such that for a linear/non-linear feature mapping \( F : \mathbb{R}^N \to \mathbb{R}^{N_r} \) we have

\[
\sum_{m,l \in \mathcal{I}, m \neq l} \left| \left| F(Y_m) - F(Y_l) \right| \right|^2 < \epsilon
\]

(4)

where mapping \( F \) is determined according to different applications. The complete joint diagonalization problem can be formulated by the following optimization problem:

\[
\min_{Y_1,\ldots,Y_M} \sum_{m \in \mathcal{I}} ||Y_m^T L_m Y_m - \Lambda_m||_F^2 + \mu \sum_{m,l \in \mathcal{I}} \left| \left| F(Y_m) - F(Y_l) \right| \right|^2
\]

(5)

The optimized results \( Y_1^*, \ldots, Y_M^* \) are the demanded synchronized spectral basis. The resulting vectors not only have the same properties as ordinary eigenvectors of image graph \( \mathcal{G}_1, \ldots, \mathcal{G}_M \) respectively, but are also matched in pairwise fashion under feature transform \( F \).

In practice, each quasi-eigenvector \( y_{m,k} \) can be considered as the linear combination of \( \{\xi_k(\mathcal{G}_m)\}_{k=1}^K \). This assumption resolves the ambiguity of \( Y_m \) and makes the optimization more effective. This is done by letting \( Y_m = U_m A_m \) where \( A_m \) is a \( K \times K' \) matrix variable for \( K' \leq K \) and \( U_m = [\xi_1(\mathcal{G}_m), \ldots, \xi_K(\mathcal{G}_m)] \).

We also adopt the Fourier coupling [30] in the diagonalization, and let \( F \) as the matrix of discrete Fourier bases. \( F \) is now a \( N' \times N \) matrix which constitutes \( N' \) vectorized discrete 2D Fourier bases. Let \( \tilde{\Lambda}_m = \text{diag}(\lambda_1(\mathcal{G}_m), \ldots, \lambda_K(\mathcal{G}_m)) \), the optimization problem (5) is then modified as:

\[
\min_{A_1,\ldots,A_M} \sum_{m \in \mathcal{I}} ||A_m^T A_m - \tilde{\Lambda}_m||_F^2 + \mu \sum_{m,l \in \mathcal{I}} \left| \left| F(U_m A_m) - F(U_l A_l) \right| \right|^2
\]

subject to: \( A_m^T A_m = \text{Id} \) for all \( m \in \mathcal{I} \)

(6)

Figure 3 shows an example of spectral synchronization using (6) over a set of images with different sizes, colors/textures.

Figure 3 shows an example of spectral synchronization using (6) over a set of images with different sizes, colors/textures.
Figure 3: Example of synchronized spectral bases (left) v.s. unsynchronized spectral bases (right) using the same Laplacian matrix. Each column of synchronized bases are considered in the same spectral channel and are forced to be correlated by equation (6).

4.3. Synchronized Superpixels

In SSN, the image part-part correspondences are represented as synchronized superpixels (SSP). The corresponding regions across different images will be assigned the same superpixel label. The SSPs naturally provides an implicit registration mapping from one image to another. The overall spectral synchronization and superpixel labeling is illustrated in Figure 2 with the details explained as follows.

**Single-scale SSP.** The synchronized superpixels are immediately available from the synchronized spectral bases. The superpixel label assignment is similar to the classical spectral segmentation [32] where the spectral bases are clustered by K-means to obtain $K$ superpixels of the image. Unlike the classical approach, we conduct the K-means clustering across images for the synchronized spectral bases. The resulting superpixels thus have cross-image labeling, which represent the part-part correspondences described by the synchronized spectral bases. The generation of single-scale SSP can be summarized as Algorithm 1.

Once the synchronized superpixels are generated, for the $n$th superpixel in all images $\{S_{1,n}, \ldots, S_{M,n}\}$ are correlated and assigned to the same label. The corresponding regions of the superpixels in $I_1, \ldots, I_M$ are matched accordingly. For images with simple foreground/background structures, i.e., image with uniform background colors/textures, the single-scale SSP is sufficient for extracting the foreground objects and their cross-image correspondences.
Algorithm 1: Synchronized Superpixels Generation

Input: # of SSP $n$, image set $\{I_1, \ldots, I_M\}$

Output: Superpixels $\{S_{m,1}, S_{m,2}, \ldots, S_{m,n}\}_{m=1}^M$

1. Construct spectral graph $G_m$ for each $I_m$;
2. Set $\Lambda_m = \text{diag}(\lambda_1(G_m), \ldots, \lambda_K(G_m))$; $\tilde{\Lambda}_m = \text{diag}(\lambda_1(G_m), \ldots, \lambda_{K'}(G_m))$; $U_m = \begin{bmatrix} \xi_1(G_m) & \cdots & \xi_K(G_m) \end{bmatrix}$ for all $m = 1, \ldots, M$;
3. Compute optimization (6) using $\{\Lambda_m, \tilde{\Lambda}_m, U_m\}_{m=1}^M$, obtaining $Y_m \leftarrow U_m A_m$ for $m = 1, \ldots, M$;
4. Stack matrices $Y \leftarrow [Y_1 Y_2 \cdots Y_M]$;
5. Cluster rows of $Y$ into $n$ clusters using K-means;
6. For all $m = 1, \ldots, M$, partition $I_m$ to $\{S_{m,1}, \ldots, S_{m,n}\}$ according to the row clusters of $Y$;

Multi-scale SSP. The spectral synchronization and superpixel partitioning can be iteratively applied, generating multiscale synchronized superpixels. As shown in Figure 2, we first decompose and synchronize the spectral bases of the input images (Scale 1), generating the set of synchronized spectral bases and subsequently the synchronized superpixels. For each synchronized superpixel $S$, the corresponding image regions forms a set of subgraphs out of the original input images. We then apply the spectral decomposition and synchronization to these subgraphs, obtaining a new set of synchronized superpixels (Scale 2). The sub-decomposition can be carried out to even finer scales (Scale 3, 4, etc.) to obtain smaller synchronized superpixels for refined part-part correspondences. At each scale, the single-scale SSP algorithm can be directly used for the generation of the SSP at that scale. In this situation, the input of the single-scale SSP algorithm will be the masked image subregions instead of the whole images, as shown in Figure 2.

Using the multi-scale SSP, it is easy to identify the large shape distortion across images by simply checking the size changes of the SSPs. For the SSPs extracted from a set of images, if a particular SSP has more singular shrink-
age/dilatation than other SSPs in a particular image, then this image will be considered as an outlier which doesn’t share the common shape with other images. We apply this rule in identifying the homogenous images to start the shape discovery.

5. Shape Discovery and Segmentation

The part-part image correspondences obtained by synchronized superpixels significantly simplifies the detection and segmentation of common shapes in the synchronized images. If a shape pattern is recurring in the image set, it will be formed by a subset of synchronized superpixels. Thus the spatial connectivity and relative locations of the subset of SSPs can be enhanced and piled up according to their repetitions in the images. The precise object shapes can then be extracted by the guidance of the recurring SSP patterns.

5.1. Shape Discovery via Synchronized Superpixels

The accumulation and grouping of SSPs is implemented by a modified common sketch algorithm [5], which generates a SSP map of foreground likelihood. Intuitively, the common sketch algorithm do the following accumulation: if the connectivity (similarity) between two SSP remain invariant across their counterparts in all images, they will be part of the common shapes area. Otherwise, their connectivity is vulnerable and is likely to be separated into foreground/background.

Common Shape Sketches over Synchronized Superpixels. The computation of common shapes is done by a quadratic programming optimization. For each pair of superpixels in the same image, we assign their similarity by a weight in \([-1, 1]\). The weight close to 1 indicates the superpixels are highly similar (an attraction), while the weight close to \(-1\) indicates the superpixels are highly distinct (a repulsion). As the superpixels are already synchronized by SSN, each image \(I\) can construct the affinity matrix of superpixels \(\tilde{W}\), with the elements showing the similarity of distinction between SSP pairs. The matrix
\( \tilde{W} \) is defined as
\[
\tilde{W}(i,j) = \frac{1}{2}(R(S_i, S_j) + R(S_j, S_i))e^{-d_{HD}(S_i, S_j)/\sigma}
\]  
where \( R(\cdot, \cdot) \) is the similarity measure between the superpixel \( S_i \) and \( S_j \) and \( d_{HD} \) is the Hausdorff distance of the superpixels with constant \( \sigma > 0 \). In practice, \( R(\cdot, \cdot) \) is implemented as the sum of square distance (SSD) for \( S_i \) and \( S_j \) where the distance is normalized to \([-1,1]\) accordingly to represent the attraction and repulsion between them.

For each synchronized superpixel \( S \), we seek the binary sketch map \( s : S \rightarrow \{-1, 1\} \) such that \( S \) will be a foreground superpixel if \( s = 1 \), and otherwise a background superpixel if \( s = -1 \). Suppose each image in \( \{I_1, \ldots, I_M\} \) generates \( n \) synchronized superpixels and matrix \( \tilde{W}_m \), we construct a \( n \times n \) Laplacian
matrix $\tilde{L} = \tilde{D} - \frac{1}{M} \sum \tilde{W}_m$ where $\tilde{D}_m$ is the diagonal matrix with $\tilde{D}(i,i) = \frac{1}{M} \sum_{j,m} \tilde{W}_m(i,j)$. The computation of $s$ is conducted by the optimization:

$$\min_s s^T \tilde{L} s, \quad \text{subject to } -1 \leq s \leq 1. \quad (8)$$

The resulting $s$ maps the SSPs of each image to a sketch map of fore/background ranges in $[-1, 1]$. The common shapes can be obtained by thresholding the maps.

**Segmentation from Shape Sketches.** Using the sketch maps, we first apply the Otsu’s thresholding for each sketch map to extract the binary shape regions. The binary images obtained can then serve as the initial guess of the actual object shapes. We next apply the Grabcut [29] iteratively to the image, using the corresponding binary shape as the foreground input. The foreground is iteratively updated as the GrabCut result at each loop until the segmentation is stabilized. The final shapes are then obtained.

An example of the common shape extraction is shown in Figure 4. The images are first correlated and decomposed into synchronized superpixels by Algorithm 1, then are transformed to the confidence maps using optimization (8). The resulting confidence maps simultaneously outline the common shapes in all images. With a simple thresholding operation, the final shape segmentations can be easily obtained.

A useful property of our method is that the segmentation can handle different image modalities. As demonstrated in Figure 5, the sketch images which only contain contours can be jointly segmented with the natural images. The synchronization automatically correlates the corresponding parts across the two type of images, making the shape discovery more versatile in real images.

**5.2. Iterative Superpixels Transfer**

In practice, the common shapes in images can have large scale differences or with significant nonrigid/rigid deformations. For this case, we can start with SSN synchronization for a subset of roughly aligned images to construct an initial set of SSP, then iteratively transfer the SSP to images with large scale appearance changes or deformations. The iterative SSP transfer is implemented...
Our method can directly handle both modalities and extract correspondences among the images.

**SSN-based Contour Detectors.** As shown in Figure 6, for the set of roughly aligned images correlated by SSN, one can obtain a set of contour fragments by directly extracting edges in the synchronized spectral bases. These ‘principal contours’ serve as detectors for perspective shapes in new input images. The detectors are applied following the chamber matching algorithm in [33]. The advantage of using the contours extracted from synchronized spectral bases is that the contours are significant as spectral bases remove most noises of the images. Moreover, all extracted contours are correlated by SSN, such that they can be compensated for each other. The contours then serve as shape detectors for detecting common shapes in multi-scales and deformation.

**Superpixel Transfer via SSN Re-synchronization.** The contour fragments detectors identify the location of the common shapes in arbitrary new input image. Sub-image around the detected shapes can be extracted and sent to a new spectral synchronization with the initial spectral bases. Similar to (5), for the extracted sub-image $I_w$, we construct Laplacian $\mathcal{L}_w$ the re-synchronized...
spectral bases $Y_w$ can be obtained by optimization:

$$
\min_{Y_w} \|Y_w^T L_w Y_w - \tilde{\Lambda}_w\|^2_F + \mu \sum_{m \in I} \|F(Y_m) - F(Y_w)\|^2
$$

(9)

where $I = \{I_1, \ldots, I_M\}$ is the set of SSN correlated images. In other words, $I$ serves as the training set for supervising the spectral synchronization of $I_w$.

The resulting $Y_w$ can be clustered to a new set of superpixels which transfer the SSP of image set $I$ to $I_w$.

6. Experiments

Our unsupervised shape discovery method is evaluated on the three datasets (ETHZ Shape, Caltech-256, Internet Images [19]) with varying difficulty. We present the qualitative and quantitative results on the above datasets with comparisons of state-of-the-art methods. Two standard evaluation metrics: precision $P$ (ratio of correctly labeled pixels) and Jaccard similarity $J$ (intersection over the results and ground truth segmentations), are used in the quantitative evaluations. The Jaccard similarity is defined as:

$$
J(A, B) = \frac{|A \cap B|}{|A \cup B|}
$$

(10)

for the region $A$ (i.e., the resulting shape region) and $B$ (i.e., the ground-truth).
Figure 7: The example of shape discovery using superpixel transferring. (a) The shapes detected by the contour detectors in Figure 6. (b) The new superpixels obtained by the resynchronization via optimization (9), using the original superpixels in Figure 4. (c) Our final segmentations obtained by Grabcut with common sketches as the prior. (d) The state-of-the-art results of [23]. Our results can better preserve the shape completeness than [23].

We collect a number of shape discovery and segmentation methods [34] [35] [25] [36] [19] [23] [29] for comparison. For some of the methods, a pre-training or coarse labeling were needed which cropped and resized the images into object-centered ones. We list the specific requirements of these methods in Table 1. Our method, as well as [25] and [19], can be directly applied on the raw images. Also, our and [19] can automatically discard the unmatched noise image to form a homogenous seed image set.

According to the algorithm setting and labeling requirements of the testing methods, we conduct the tests and comparison in two phases. In phase 1 (Ph1), we test direct unsupervised shape discovery (Step 1 to Step 2 in Figure 1) in ‘homogenous images’: the images that have common shapes around the center with roughly similar sizes and poses. In phase 2 (Ph2), we test the seed-based shape discovery (Step 1 to Step 3 in Figure 1) on ‘incoherent images’: images
TABLE 1: Comparison of the pre-training or labeling requirements of the testing methods. Our method is unsupervised and does not require labeled data for pre-training.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Initial training/labling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>N/A</td>
</tr>
<tr>
<td>BoSS [34]</td>
<td>Hand-drawn object outlines</td>
</tr>
<tr>
<td>ClassCut [35]</td>
<td>Objectness frames [37]</td>
</tr>
<tr>
<td>CoSand [25]</td>
<td>N/A</td>
</tr>
<tr>
<td>Model-driven [36]</td>
<td>Hand-drawn bounding-boxes</td>
</tr>
<tr>
<td>Discover &amp; Seg [19]</td>
<td>N/A</td>
</tr>
<tr>
<td>Visual Knowledge [23]</td>
<td>Exemplar-based Detection [38]</td>
</tr>
<tr>
<td>GrabCut [29]</td>
<td>Hand-drawn bounding-boxes</td>
</tr>
</tbody>
</table>

with common shapes under different sizes and poses. The initial seed images are manually selected from the test datasets for building SNN-based shape detectors. The split of Ph1 and Ph2 is for a more fair comparison of the performance between the fully unsupervised methods and the weakly supervised methods, as well as their performance on small homogenous dataset and large scale heterogeneous dataset respectively.

6.1. Phase 1: Tests on Direct Shape Discovery

We compare our method to the unsupervised segmentation methods BoSS [34], ClassCut [35], Model-driven [36], CoSand [25], and GrabCut [29]. Three datasets are used in this phase: ETHZ, Caltech 101, and Weizmann Horse. The GrabCut method is considered as the baseline. The qualitative results of the test are shown in Figure 8, Figure 9, Figure 10 and Figure 11 while the quantitative results are reported in Table 2 and Table 3. Particularly, Table 2 shows the performance of the coarse labeling method and Table 3 particularly presents the fully unsupervised methods.

During the test, 10 homogeneous images from each class (50 in total) are randomly selected for the evaluation. The baseline results are obtained by taking the GrabCut over the test image with a manually drawn object bounding...
Figure 8: Example results on coherent images in ETHZ Shapes dataset. From top to bottom: original inputs, synchronized superpixels for each shape classes, the final shape segmentations.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Apple</th>
<th>Bottles</th>
<th>Giraffes</th>
<th>Mugs</th>
<th>Swan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (Ph1)</td>
<td>98.1</td>
<td>97.9</td>
<td>95.5</td>
<td>96.5</td>
<td>96.8</td>
</tr>
<tr>
<td>BoSS [34]</td>
<td>98.4</td>
<td>97.3</td>
<td>94.1</td>
<td>96.4</td>
<td>95.1</td>
</tr>
<tr>
<td>Model-driven [36]</td>
<td>96.3</td>
<td>87.7</td>
<td>82.3</td>
<td>88.5</td>
<td>92.7</td>
</tr>
<tr>
<td>Baseline (GrabCut)</td>
<td>95.9</td>
<td>85.3</td>
<td>82.1</td>
<td>89.4</td>
<td>85.0</td>
</tr>
</tbody>
</table>

Table 2: Comparative results of correct pixel percentages (P) over the coherent images in ETHZ dataset. The average precision is reported ([36] and ours) due to the random training/testing splitting.

As shown in Table 2 and Table 3, our method has significant better performance than the other methods. Note that segmentation method of [35] still involve weakly supervised training for shape prior. Our method does not require this prior and can be computed in fully unsupervised style. Because of the adaptive common shape sketches obtained from SSN, we can obtain a better initial guess of the common shape areas and thus obtain the better segmentation results than the other methods, even the similar GrabCut mechanism is employed in those methods. Moreover, as shown in the qualitative results, our method naturally provide the part-part correspondences between images while
Table 3: Comparative result of correct pixel percentages over the coherent/incoherent images in Weizmann Horse (W) and Caltech (C) datasets. The average results of our Phase 1 (Ph1) test reported topped the performance in all classes.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Horse (W)</th>
<th>Face (C)</th>
<th>Plane (C)</th>
<th>Motor (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (Ph1)</td>
<td>90.1</td>
<td>93.2</td>
<td>90.2</td>
<td>92.8</td>
</tr>
<tr>
<td>CoSand [25]</td>
<td>84.3</td>
<td>90.1</td>
<td>73.8</td>
<td>87.6</td>
</tr>
<tr>
<td>ClassCut [35]</td>
<td>86.2</td>
<td>89.0</td>
<td>89.8</td>
<td>90.3</td>
</tr>
<tr>
<td>Baseline (GrabCut)</td>
<td>80.4</td>
<td>87.3</td>
<td>84.1</td>
<td>85.1</td>
</tr>
</tbody>
</table>

Other methods only have the segmented regions as the results (i.e., [35]). The correspondences between subparts can be further utilized in extraction of the common shape subparts for subcategory discovery (i.e., [23]).

6.2. Phase 2: Tests on Shape Discovery with Unsupervisedly-Trained Seeds

Ph2 is for the test of large scale datasets where most of Ph1 testing methods will fail. In this test, we perform our shape discovery over the incoherent images in the internet image dataset ([19]) with comparison to the methods of [39], [25], [19], and [23]. Our method perform successful shape discovery in the testing datasets, and the comparative results show our method have better shape extraction accuracy and more robust foreground/background discrimination.

The dataset contains three main object categories: car, horse, airplane with various object scales and poses. During the test, 10 seed images in one categories are first selected to build up the initial SSN. The resulting synchronized
Figure 10: Example results on direct/seed-based shape discovery in Face and Airplane class in Caltech dataset. The face example requires the first three images as the seed images, while the airplane example does not require seeds to obtain the final segmentation.

superpixels from these images are then transfer to other images in the set to obtain the segmentation on them. The overall qualitative results of the test are shown in Figure 9, Figure 10, and Figure 11 while the quantitative results are in Table 4. Note that the testing method [23] contain the extra preprocessing methods [37] for object localization. We consider the unsupervised discovery in phase 1 as the preprocessing of phase 2 test. The segmentation will propagate from a small set of homogeneous image subset (seeds) to all other images and evaluation will be conducted on both the test images and the seed images. We add the Ph1 results of our seed images for the complete comparison of accuracy lost/gain in the two stages.

As the iterative SSP transfer algorithm contains an additional shape detection step, we first evaluate the performance of the SSN-based shape detectors then the final segmentation. The construction and representation of our shape
Figure 11: Example results on direct/seed-based shape discovery in the internet image dataset. Two groups of directly discovered shapes (left) are acted as double seeds in discovering the shapes on the multi-object images (right).

Figure 12: Additional comparative examples of shape discovery using superpixel transferring, compared to the results of Car dataset in [23]. Our method is better in keeping the common object shape outlook in all images.

detector is similar to those in [40] and [34], thus we compare our detection performance with these two methods. The detection results are presented in Figure 15 using the form of precision-recall curve. Our unsupervisedly generated shape detectors are competitive to [40] and [34]. This is because the synchronized spectral bases, as their counterpart in the unsynchronized cases, preserve the salient contour fragments while removing the other small edge segments in the images. The SSN-based shape detectors can capture the most salient shape parts of the objects, improving the accuracy and robustness of the shape detection.

We then evaluate the performance of the final segmentation results obtained
Figure 13: Additional comparative examples of our method and the results of Horse dataset in [23]. Our method is better in discriminating the background and foreground.

Figure 14: Additional comparative examples of our method and the results of Horse dataset in [23]. Our method is better in preserving the continuous details of the shapes.

from the iterative SSP transfer. As shown in Table 3, our method has better final segmentation than the other four methods. In fact, all testing methods including ours use the same GrabCut algorithm as the last step. Our method can outperform the other three because our shape mask input for the GrabCut is image driven which is highly adjustable to the image object. The synchronized superpixels of each image adaptively adhere to the object boundaries while the shape masks used in [19] and [23] can only loosely matched the image objects. The loose masks make the Grabcut segmentation miss some important portions of the objects.

Illustrative comparative examples are presented in Figure 12, Figure 13, and Figure 14. These examples particularly demonstrate how our method perform in the preservation of overall shape outlook, preservation of small shape features/parts, and the discrimination under complicated foreground/background.
Table 4: The comparative results of our method with the state-of-the-arts in the internet images dataset [19]. The pixel precision (P) and Jaccard similarity (J) are reported. Our initial segmentation result (Ph1) for the seed images are also presented to show the changes when large image collections are involved.

<table>
<thead>
<tr>
<th></th>
<th>Car</th>
<th>Horse</th>
<th>Airplane</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>J</td>
<td>P</td>
</tr>
<tr>
<td>Ours (Ph1, seed images)</td>
<td>89.1</td>
<td>70.1</td>
<td>88.4</td>
</tr>
<tr>
<td>Ours (Ph2)</td>
<td>87.1</td>
<td>68.1</td>
<td>89.2</td>
</tr>
<tr>
<td>Visual Knowledge [23]</td>
<td><strong>87.65</strong></td>
<td>64.86</td>
<td>86.16</td>
</tr>
<tr>
<td>Discover &amp; Seg [19]</td>
<td>85.38</td>
<td>64.42</td>
<td>82.81</td>
</tr>
<tr>
<td>CoSand [25]</td>
<td>68.85</td>
<td>0.04</td>
<td>75.12</td>
</tr>
<tr>
<td>MutiClass Coseg [39]</td>
<td>59.20</td>
<td>35.15</td>
<td>64.22</td>
</tr>
<tr>
<td>Baseline (GrabCut)</td>
<td>54.5</td>
<td>40.0</td>
<td>45.9</td>
</tr>
</tbody>
</table>

Figure 15: The precision-recall detection performance of our SSN-based detectors (Sec. 5.2) in three typical shape classes from Caltech-101 dataset. Our detectors are competitive to the supervised approaches in [40] and [34].
As the figures show, the false segmentations in [23] are caused by the high similarity of the foreground colors and the background colors. The occasional missing of the objects in detection will also cause the mis-segmentation. Instead, our synchronization method provides strong correspondence across images while it still keeps the local area continuity under the spectral graph model. This avoids the missing parts and mis-segmentation in shape extraction.

7. Discussion

Compared with other testing methods, our method is a groupwise method with linear complexity $O(N)$. The linear complexity is similar to other groupwise method like CoSand [25] and Many-to-one [15]. The complexity of pairwise methods such as the joint discovery and segmentation [18] and [19] will be $O(N^2)$, as pairwise image matching is needed in this model. Other testing methods are complicated combination of multiple algorithms thus their exact complexity will be arguable. We consider their complexities lie in the range between $O(N)$ and $O(N^2)$. Also, the proposed method used the same synchronization model in the seed segmentation and the test image segmentation. This is unlike the other testing method (i.e., [23]) where an extra independent preprocessing algorithm is needed. The overall complexity of our method is thus lower than the other hyper methods.

8. Conclusions

We proposed an unsupervised shape discovery using the part-part correspondences across images. The images are represented as spectral graphs and their correspondences are identified by a novel Spectral Synchronization Network model. The network correlates different images by synchronizing their spectral bases, such that multiple images with different colors/textures/modalities can be matched simultaneously. The cross-image correspondences provide natural shared shape features for groupwise shape analysis. Using the image correspondences, we build up the synchronized superpixels and extract the common
shape patterns via common shape sketch algorithm. The resulting segmentation guided by the extracted synchronized superpixels shows better accuracy than the existing unsupervised discovery methods. The proposed method will largely alleviate the need of labeling in existing supervised or weakly supervised approaches, enabling a fully unsupervised image understanding for the object shapes.

References


Yunliang Cai is currently a Research Associate in Dartmouth College, Hanover NH, USA. He obtained his Ph.D in computer science from Hong Kong Polytechnic University. He was a Postdoctoral Fellow in the Digital Imaging Group of London (DIG) and the Department of Medical Biophysics, Western University, London ON, Canada. His research interest includes medical image analysis, image-guided surgery, and computer vision.

Dr. Shuo Li is an associate professor in department of medical imaging and medical biophysics in the University of Western Ontario and scientist in Lawson Health Research Institute. Before this position he was research scientist and project manager in general electric (GE) healthcare, Canada for 9 years. He fund and direct the Digital Imaging Group of London (http://digitalimaginggroup.ca/) since 2006, which is a very dynamic and highly multiple disciplinary collaboration group. He received his Ph.D. degree in computer science from Concordia University 2006, where his PhD thesis won the doctoral prize giving to the most deserving graduating student in the faculty of engineering and computer science. He has published over 100 publications; He is the recipient of several GE internal awards; He serves as guest editors and associate editor in several prestigious journals in the field; He servers as program committee members in highly influential conferences; He is the editors of five springer books. His current interest is development intelligent analytic tools to help physicians and hospital administrative to handle the big medical data, centered with medical images.