Representative Discovery of Structure Cues for Weakly-Supervised Image Segmentation

Luming Zhang, Yue Gao, Yingjie Xia, Ke Lu, Jialie Shen, and Rongrong Ji

***Abstract*—Weakly-supervised image segmentation is a challenging problem with multidisciplinary applications in multimedia content analysis and beyond. It aims to segment an image by leveraging its image-level semantics (i.e., tags). This paper presents a weakly-supervised image segmentation algorithm that learns the distribution of spatially structural superpixel sets from image-level labels. More specifically, we first extract graphlets from a given image, which are small-sized graphs consisting of superpixels and encapsulating their spatial structure. Then, an efficient manifold embedding algorithm is proposed to transfer labels from training images into graphlets. It is further observed that there are numerous redundant graphlets that are not discriminative to semantic categories, which are abandoned by a graphlet selection scheme as they make no contribution to the subsequent segmentation. Thereafter, we use a Gaussian mixture model (GMM) to learn the distribution of the selected post-embedding graphlets (i.e., vectors output from the graphlet embedding). Finally, we propose an image segmentation algorithm, termed representative graphlet cut, which leverages the learned GMM prior to measure the structure homogeneity of a test image. Experimental results show that the proposed approach outperforms state-of-the-art weakly-supervised image segmentation methods, on five popular segmentation data sets. Besides, our approach performs competitively to the fully-supervised segmentation models.**

***Index Terms*—Structure cues, graphlet, weakly supervised, segmentation, active learning.**

I. Introduction

As a preliminary step, image segmentation has been widely used in many multimedia applications, e.g., image cropping [30], photo aesthetics ranking [2], [15], and scene parsing [9]. Typically, these applications require the images to be ideally segmented, i.e., each segmented region corresponds to a completed semantic component. Nevertheless, targeting at an optimal segmentation needs extensive human supervision, as automatic approaches are far from satisfactory. However, most existing applications are built upon unsupervised image segmentation methods, whose performance is unsatisfactory due to the lack of high-level cues. For example, many segmented regions partially cover one or multiple semantic objects, which largely degenerate the segmentation quality to conduct the subsequent application scenarios.

Inspired by the idea of supervised image retrieval [12], [25], [26], image-level labels are cheaply available, i.e., can be efficiently and accurately acquired. While such supervision is not specific to any regions, is it possible to make use of such "weak supervision" to facilitate image segmentation, that is, image-level labels, to improve image segmentation. In this paper, weakly-supervised image segmentation is defined as: in the training stage, semantic labels are only at the image level, without regard to their specific object/scene location within the image. Given a test image, the goal is to predict the semantic labels to every pixel. However, weakly-supervised image segmentation is a challenging problem due to two factors:

- The intrinsic ambiguity of image-level labels: compared with the pixel-level labels used in fully-supervised segmentation models, image-level labels are much coarser cues which are difficult to be incorporated into the segmentation model.
- The spatial structure is neglected in measuring the homogeneity of superpixels[1]: beyond the appearance features, the spatial structure of superpixels is also important for measuring their homogeneity, which is however not taken into consideration in the existing segmentation models [7], [8]. As shown in Fig. 1, the yellow pyramid and the sand have similar superpixel appearances. However, compared to the sand region, the pyramid superpixels feature in their unique triangular patterns, thus they should be assigned with strong homogeneity and encouraged to merge.

To address the above two problems, we propose to learn the distribution of graphlets from image-level labels, which are then used to guide the image segmentation process. To capture the spatial structure of superpixels, we extract graphlets by connecting spatially neighboring superpixels. Herein, graphlets are small-sized graphs that capture the neighboring structures of superpixels. Considering that graphlets of different sizes are in-

Manuscript received May 09, 2013; revised August 18, 2013 and October 15, 2013; accepted October 15, 2013. Date of publication November 28, 2013; date of current version January 15, 2014. This work was supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office. This work was also supported by the Nature Science Foundation of China (No. 61373076), the Fundamental Research Funds for the Central Universities (No. 2013121026), and the 985 Project of Xiamen University. Natural Science Foundation of Chinaunder grant number 61002009, Key Science and Technology Program of Zhejiang Province of Chinaunder grant number 2012C01035-1, and Zhejiang Provincial Natural Science Foundation of Chinaunder grant number LZ13F020004 (Corresponding author: Y. Gao). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Cees Snoek.

L. Zhang and Y. Gao are with the School of Computing, National University of Singapore, Singapore (e-mail: kevin.gaoy@gmail.com).

Y. Xia is with the Hangzhou Institute of Service Engineering, Hangzhou Normal University, Zhejiang, China.

L. Ke is with the Graduate University of Chinese Academy of Sciences, Beijing 10049, China.

J. Shen is with the School of Information Systems Singapore Management University, Singapore.

R. Ji is with the Department of Cognitive Science, Xiamen University, Xiamen, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2293424

[1]In the image segmentation community, the homogeneity of superpixels is a basic and fundamental concept which reflects the probability of pairwise or multiple superpixels sharing a common semantic label.

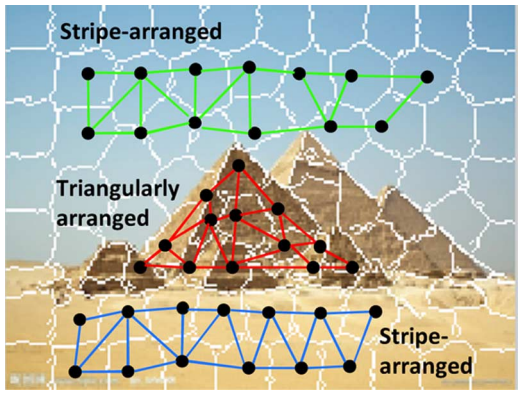

Fig. 1. The superpixel mosaic of an example image.

comparable in the Euclidean space, we project graphlets onto the Grassmann manifold, based on which a manifold embedding incorporates image-level labels into graphlets. Through the embedding, different-sized graphlets are transformed into equal-lengthed feature vectors. Noticeably, there are a large number of graphlets in an image given the graphlet size is moderate. And not all graphlets are representative for the semantic category, i.e., many of them are redundant for segmentation. Therefore, a selection algorithm is developed to discover a few semantically representative graphlets from an image, which are further used to learn the distribution of graphlets. Given the learned graphlet distribution as a hint for the spatial structure of superpixels, we propose a new segmentation algorithm, called representative graphlet cut, that leverages the learned graphlet distribution.

A preliminary conference version of this work has been accepted by CVPR 2013 [31]. Our new improvement comparing to the conference version paper lies in the following aspects. First, all graphlets are employed as the homogeneity measure in [31]. As we stated, those graphlets that are not representative to a semantic category will affect the segmentation process. Inspired by the selectively viewing mechanism in human vision [32], we design a selection algorithm to acquire a few semantically representative graphlets from an image as the homogeneity measure. Thereby, a more efficient and effective segmentation framework is achieved, which has been demonstrated by the experimental results. Second, in the experiment, compared with the conference paper, more extensive experiments are conducted on additional data sets: MSRC-21 [18], VOC 2008–2010 [4]. More importantly, we observe that by introducing the representative graphlet selection strategy, an increase of 6%–8% segmentation accuracy is achieved on each of the five experimental data sets.

II. Related Work

Recently, several weakly-supervised image segmentation methods [20]–[24] have been proposed, focusing on developing statistical models to transfer image-level labels into superpixels unary or pairwise potentials[2]. Verbeek *et al.* [20] proposed an aspect model to estimate pixel-level labels for each image, which is modeled as a mixture of latent topics. Vezhnevets *et al.* [21] formulated weakly-supervised image segmentation as a multiple instances learning problem. However, the unary potential used in [20] and [21] fail to model the interactions between superpixels, which are important for smoothing superpixel labels. To model the relationships among superpixels, Verzhnevets *et al.* [22] proposed a graphical model, termed multi-image model (MIM), that integrates image appearance features, image-level labels and superpixel labels into one single network. To refine the MIM-based segmentation, Verzhnevets *et al.* [23] designed an active learning scheme [29] to select superpixels that are semantically most uncertain within an image. The selected superpixels are accurately labeled by querying an oracle database, and they guide the label inference for the remaining superpixels. Moreover, Verzhnevets *et al.* [24] developed a parametric family of structured models, where multi-channel visual features were employed to form the pairwise potential, and the weights of each channel is computed by minimizing the discrepancy between superpixels labeled by differently-trained models.

One weakness of these weakly-supervised segmentation methods is the low descriptive unary/pairwise potentials, resulting in many ambiguous segment boundaries. To alleviate this problem, high-order potentials are exploited that measures the probability of multiple superpixels belonging to a semnatic label. Kohli *et al.* [8] proposed a high-order conditional random field for image segmentation, where the high-order potentials are defined over pixel sets. In [17], Rital *et al.* generalized the conventional normalized cut into hypergraph cut, where each hyperedge connects multiple spatially neighboring superpixels. However, hypergraph cut has two limitations: 1) supervision incorporation is difficult, and 2) label inference is computationally inefficient. To overcome these limitations, Kim *et al.* [7] developed a supervised high-order correlation clustering technique for image segmentation. Based on the structured support vector machine and the linear programming relaxation, both the parameter learning and segmentation process are carried out efficiently. Notably, these approaches are either unsupervised or fully-supervised, and it is difficult to transform them into a weakly-supervised version. Moreover, the spatial structure of superpixels is neglected.

III. The Proposed Approach

A. Graphlet Extraction and Representation

Superpixel is a commonsense practice in efficient image segmentation to deal with the gigantic amount of pixels. In our implementation, the superpixels are generated based on simple linear iterative clustering (SLIC) [1]. After that, graph based approach such as Region Adjacency Graph (RAG) is typically leveraged to model the spatial adjacency among superpixles, i.e.,

$$\mathcal{G} = (V, E), \tag{1}$$

where V is a set of vertices, each representing a superpixel; E is a set of edges, each connecting pairwise spatially adjacent superpixels.

An image usually contains multiple semantic components, each spanning several superpixels. Given a superpixel set, two observations can be made. First, the appearance and

[2]In the image segmentation community, the homogeneity of superpixels is a basic and fundamental concept which reflects the probability of pairwise or multiple superpixels sharing a common semantic label.

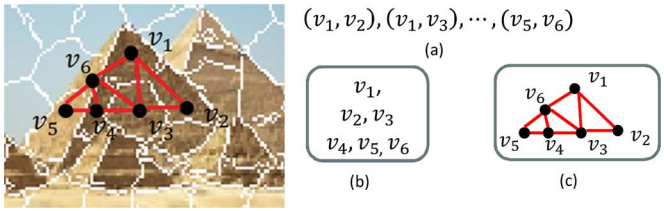

Fig. 2. Different types of superpixel homogeneity.

spatial structure of superpixels collaboratively contribute to their homogeneity. Second, the more their appearance and spatial structure correlate with a particular semantic object, the stronger their homogeneity. For instance as shown in Fig. 1, the superpixel set in the sky region and the superpixel set in the sand region have similar spatial structure but different superpixel appearance, thus they should be assigned with different homogeneities. Compared with the stripe-distributed yellow superpixels, the stripe-distributed blue superpixels (in Fig. 1) appear more common in semantic classes, such as lake and river, which indicates they are low correlated with a particular semantic object, thus should be assigned with a weaker homogeneity. On the other hand, the superpixel sets covering pyramid and sand have similar superpixel appearance but different spatial structure, thus should also be assigned with different homogeneities. Compared with the stripe-distributed yellow superpixels, the triangularly-distributed yellow superpixels are unique for the Egyptian pyramid, thus they should be assigned with a stronger homogeneity.

We propose graphlets to capture the appearance and spatial structure of superpixels. The graphlets are obtained by extracting connected subgraphs from an RAG. The size of a graphlet is defined as the number of its constituent superpixels. In this work, we restrict to study small-size graphlets because: 1) the number of all the possible graphlets is exponentially increasing with grpahlet sizes; 2) the graphlet embedding implicitly extends the homogeneity beyond single small-sized graphlets. (as shown in Section III-B); 3) empirical results show that segmentation accuracy stops increasing when the graphlet size increases from 5 to 10. That means small-sized graphlets are descriptive enough. Let T denote the maximum graphlet size, we extract graphlets of all sizes ranging from 2 to T. The graphlet extraction is based on depth-first search, which is computationally efficient. Besides, our approach is also memory-light. Given 50 superpixels in an image, and assuming the average superpixel degree is 5 and the maximum graphlet size is also 5, there are $50 * 5^5/5! + \cdots + 50 * 5^2/2! \approx 4300$ graphlets, which, after embedding, are transformed into 4300 low-dimensional feature vectors. Thus, the required storage space is very small.

It is worth emphasizing that, as a spatially structured superpixel set, graphlet-based homogeneity is a natural extension of the non-structural superpixel set homogeneity [7], [8]. As shown in Fig. 2, both the pairwise and high-order potentials represent the homogeneity of orderless superpixels, whereas the graphlet represents the homogeneity of spatially structured superpixels. If we ignore the topology encoded in graphlets, the proposed graphlet-based homogeneity reduces to the high-order superpixel homogeneity.

A quantitative description of graphlets is necessary for a computational segmentation model. Given a t-sized graphlet, we characterize the appearance of its superpixels by a matrix $\mathbf{M}_r$. Each row of $\mathbf{M}_r$ is a 137-dimensional feature vector extracted from a superpixel, i.e., a 128-dimensional histogram of gradient (HOG) [3] combined with a 9-dimenional color moment [19]. And, for the spatial structure of superpixels within a t-sized graphlet, we use a $t \times t$-sized matrix to represent it as:

$$\mathbf{M}_s(i,j) = \begin{cases} \theta(R_i, R_j) & \text{if } R_i \text{ and } R_j \text{ are spatially adjacent} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\theta(R_i, R_j)$ is the angle between the positive horizontal direction and the vector from the center of superpixel R_i to the center of superpixel R_j. Based on $\mathbf{M}_r$ and $\mathbf{M}_s$, a t-sized graphlet can be represented by a $t \times (137 + t)$ matrix, i.e.,

$$\mathbf{M} = [\mathbf{M}_r, \mathbf{M}_s]. \tag{3}$$

It can be observed that spatially neighboring graphlets in a photo are partially overlapping. This brings the property of locality of graphlets, which indicates that a graphlet and its spatially neighbors are highly correlated. Therefore, it is beneficial to exploit the local structure among graphlets when projecting them onto the semantic space. Mathematically speaking, each matrix can be deemed as a point on manifold [28], [14] and the Golub-Werman distance [27] between identical-sized matrices is:

$$d_{GW}(\mathbf{M}, \mathbf{M}') = \|\mathbf{M}_o - \mathbf{M}'_o\|_2, \tag{4}$$

where $\mathbf{M}_o$ and $\mathbf{M}'_o$ denote the orthonormal basis of $\mathbf{M}$ and $\mathbf{M}'$ respectively.

B. Manifold Graphlet Embedding

As mentioned in Section III-A, the appearance and spatial structures of semantically-consistent superpixels reflect strong homogeneity. Thus, it is necessary to integrate category information into graphlets in measuring the homogeneity of superpixels. To this end, a manifold embedding algorithm is proposed to encode image-level labels into graphlets. Besides image-level labels, two supplementary cues: image global spatial layout and geometric context, are also incorporated.

To incorporate the global spatial layout information, we enforce our embedding scheme to maximally preserve the relative distances between the graphlets. This is helpful to expand the homogeneity of superpixels across individual graphlets. Such preservation implicitly extends the homogeneity beyond the individual small-sized graphlets.

As demonstrated by Vezhnevets *et al.* [21], rough geometric context [6] effectively complements image-level labels for image segmentation. Here, rough geometric context means categorizing each pixel in an image into ground, differently oriented vertical surfaces, non-planar solid, or porous. This motivates us to integrate geometric context information into the embedding process. Intuitively, a graphlet with consistent geometric context should reflect stronger homogeneity. As shown in the right of Fig. 3, graphlet G_1 has more consistent geometric context than graphlet G_2, thus superpixels within

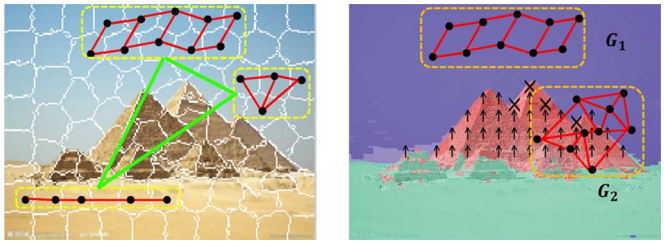

Fig. 3. Left: example of preserving global spatial layout; Right: adding rough geometric context into graphlets, ground(green), sky(blue), different oriented vertical regions(red), non-planar solid('x').

G_1 should be assigned with stronger homogeneity than those within G_2.

To capture the above three cues, namely, image-level labels, global spatial layout, and geometric context, we propose a manifold embedding algorithm with the objective function defined as (5), shown at the bottom of the page, where $\mathbf{Y} = [y_1, y_2, \cdots, y_N]$ is the row vector containing the post-embedding graphlets. The first term $\sum_{ij}[d_{GW}(\mathbf{M}_i^h, \mathbf{M}_j^h) - d_E(y_i^h, y_j^h)]^2$ describes the discrepancy between pairwise graphlet distances on the Grassmann manifold and those in the Euclidean space. The minimization of this term will maximally preserve the global spatial arrangement of the graphlets. The second term $\phi_i^h \phi_j^h$ enforces the geometric context constraint on graphlets. That is, graphlets with more consistent geometric context are assigned with larger weights. The third term $\sum_{ij} \|y_i - y_j\|^2 l_s(i,j) - \sum_{ij} \|y_i - y_j\|^2 l_d(i,j)$ encodes image-level labels into pairwise graphlets. That is, the proximity of two graphlets in feature space should be consistent with their image-level labels.

The variables in (5) are defined as follows. $\mathbf{M}_i^h$ and $\mathbf{M}_j^h$ denote two identical-sized graphlets; y_i^h and y_j^h are their low-dimensional representations; l_s is a function that measures the similarity between graphlets; l_d is a function that measures the difference between two graphlets. Let $b(G)$ denote the C-dimensional row vector containing the class label of the image corresponding to graphlets G. Denote $\vec{N} = [N^1, N^2, \cdots, N^C]^T$ a C-dimensional vector containing the number of images associated with each label, where N^c is the number of images for category c, then $l_s(i,j) = \frac{[b(G_i) \cap b(G_j)]\vec{N}}{\sum_c N^c}$ and $l_d(i,j) = \frac{[b(G_i) \oplus b(G_j)]\vec{N}}{\sum_c N^c}$. ϕ_i reflects the geometric context consistency of the i-th graphlet, which is implemented as the i-th graphlet entropy, i.e., $\phi_i = -\sum_j g_i(j) \log_2 g_i(j)$, where $g_i(j)$ is percentage of the j-th geometric context corresponding to graphlet G_i. ϕ_i^h is the geometric context obtained from the i-th graphlet in the h-th training image.

We denote $\mathbf{D}_{GW}^h = [d_{GW}(\mathbf{M}_i^h, \mathbf{M}_j^h)]$ as the matrix whose entry $d_{GW}(\mathbf{M}_i^h, \mathbf{M}_j^h)$ is the Golub-Werman distance between the i-th and j-th identical-sized graphlets extracted from the h-th image. Its inner product matrix is obtained by:

$$\tau\left(\mathbf{D}_{GW}^h\right) = -\mathbf{R}_{N_h} \mathbf{S}_{GW}^h \mathbf{R}_{N_h}/2, \tag{6}$$

where $(\mathbf{S}_{GW}^h)_{ij} = (\mathbf{D}_{GW}^h)_{ij}^2$, and $\mathbf{R}_{N_h} = \mathbf{I}_{N_h} - \vec{\mathbf{e}}_{N_h} \vec{\mathbf{e}}_{N_h}^T / N$ which is the centralization matrix. $\mathbf{I}_{N_h}$ is an $N_h \times N_h$ identity matrix, $\vec{\mathbf{e}}_{N_h} = [1, 1, \cdots, 1]^T \in \mathbb{R}^{N_h}$, N is the number of all training graphlets, and N_h the number of graphlets from the h-th training image.

Thus the first part of (5) can be rewritten as:

$$\begin{aligned}
&\arg\min_{\mathbf{Y}} \sum_h \sum_{ij} \left[d_{GW}\left(\mathbf{M}_i^h, \mathbf{M}_j^h\right) - d_E\left(y_i^h, y_j^h\right)\right]^2 * \phi_i^h \phi_j^h \\
&= \arg\min_{\mathbf{Y}} \sum_h \left\|\tau\left(\mathbf{D}_{GW}^h\right) - \tau\left(\mathbf{D}_Y^h\right)\right\|^2 * \phi_i^h \phi_j^h \\
&= \arg\max_{\mathbf{Y}} \sum_h \operatorname{tr}\left(\mathbf{Y}^h \tau(\mathbf{D}_{GW}^h \mathbf{\Phi}^h) \left(\mathbf{Y}^h\right)^T\right) \\
&= \arg\max_{\mathbf{Y}} \operatorname{tr}(\mathbf{Y} \tau(\mathbf{D}_{GW}) \mathbf{\Phi}) \mathbf{Y}^T,
\end{aligned} \tag{7}$$

where $\mathbf{\Phi} = [\phi_i, \phi_j]$ is an $N \times N$ matrix; and $\mathbf{D}_{GW}$ is a block diagonal matrix, the h-th diagonal block is $\mathbf{D}_{GW}^h$.

The second part in (5) can be rewritten into (8), shown at the bottom of the page, where $\mathbf{R} = [-\vec{\mathbf{e}}_{N-1}^T, \mathbf{I}_{N-1}]^T W_1 [-\vec{\mathbf{e}}_{N-1}^T, \mathbf{I}_{N-1}] + \cdots + [\mathbf{I}_{N-1}, -\vec{\mathbf{e}}_{N-1}^T]^T W_N [\mathbf{I}_{N-1}, \vec{\mathbf{e}}_{N-1}^T]$, and $\mathbf{W}_i$ is an $N \times N$ diagonal matrix whose h-th diagonal element is $[l_s(h,i) - l_d(h,i)] * \phi_h \phi_i$.

Based on above formulation, we can reorganize the objective function as:

$$\begin{aligned}
&\arg\max_{\mathbf{Y}} \mathbf{Y}(\tau(\mathbf{D}_{GW})\mathbf{\Phi} + \mathbf{R})\mathbf{Y}^T = \arg\max_{\mathbf{Y}} \mathbf{Y}\mathbf{Q}\mathbf{Y}^T \\
&\quad s.t.\ \mathbf{Y}\mathbf{Y}^T = \mathbf{I}_N,
\end{aligned} \tag{9}$$

$$\begin{aligned}
\arg\min_{\mathbf{Y}} & \underbrace{\sum_h \sum_{ij} \left[d_{GW}\left(\mathbf{M}_i^h, \mathbf{M}_j^h\right) - d_E\left(y_i^h, y_j^h\right)\right]^2 * \phi_i^h \phi_j^h}_{global\ spatial\ layout + geometric\ context} \\
+ & \underbrace{\left[\sum_{ij} \|y_i - y_j\|^2 l_s(i,j) - \sum_{ij} \|y_i - y_j\|^2 l_d(i,j)\right] * \phi_i \phi_j}_{image-level\ labels + geometric\ context}
\end{aligned} \tag{5}$$

$$\begin{aligned}
&\arg\max_{\mathbf{Y}} \left[\sum_{ij} \|y_i - y_j\|^2 l_d(i,j) - \sum_{ij} \|y_i - y_j\|^2 l_s(i,j)\right] * \phi_i \phi_j \\
&= \arg\max_{\mathbf{Y}} \operatorname{tr}(\mathbf{Y}\mathbf{R}\mathbf{Y}^T)
\end{aligned} \tag{8}$$

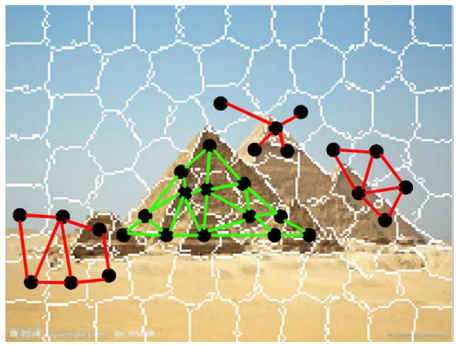

Fig. 4. An example of one highly semantics-correlated (green) and three low semantics-correlated graphlets (red).

where $\mathbf{Q} = \tau(\mathbf{D}_{GW})\mathbf{\Phi} + \mathbf{R}$ is an $N \times N$ matrix, and $\mathbf{Y}\mathbf{Y}^T = \mathbf{I}_N$ is a term to uniquely determine $\mathbf{Y}$. Note that the embedding in (9) can only handle identical-sized graphlets. Assuming the maximum graphlet size is T, the embedding is repeated T times.

C. Representative Graphlet Selection

As shown in Fig. 4, given an image, there are a large number of graphlets that are non-representative to one semantic category. Thus, homogeneity of superpixels within each of these graphlets is weak because these superpixels span different objects. These graphlets contribute little to image segmentation. Toward an efficient and effective segmentation model, it is essential to select a few representative graphlets from an image, which is achieved by a feature selection algorithm in this subsection.

An effective learning algorithm should reveal the underlying data structure. Based on the locality of graphlets in a photo, each graphlet can be linearly reconstructed by its spatial neighboring ones, where the optimal reconstruction coefficients are calculated by:

$$\begin{aligned} &\arg\min_{\mathbf{W}} \sum\nolimits_{i=1}^{N} \left\| y_i - \sum\nolimits_{j=1}^{N} \mathbf{W}_{ij} y_j \right\| \\ &s.t. \sum \mathbf{W}_{j=1}^{N} = 1, \mathbf{W}_{ij} = 0 \; if \; y_j \notin \mathcal{H}(y_i), \end{aligned} \tag{10}$$

where $\{y_1, y_2, \cdots, y_N\}$ is the post-embedding graphlets, $\mathbf{W}_{ij}$ denotes the contribution of the j-th graphlet to construct the i-th graphlet, N is the total number of graphlets in a photo, and $\mathcal{H}(y_i)$ contains the spatial neighbors of the i-th graphlets.[3]

To evaluate the representativeness of the selected graphlets, we develop a graphlet reconstruction approach. The reconstruction error reflects the quality of the selected graphlets. Let $\{z_1, z_2, \cdots, z_N\}$ be the constructed graphlets, they are determined by minimizing the following cost function:

$$\begin{aligned} &\epsilon(z_1, z_2, \cdots, z_N) \\ &= \sum\nolimits_{i=1}^{K} \|z_{s_i} - y_{s_i}\|^2 + \mu \sum\nolimits_{i=1}^{N} \left\| z_i - \sum\nolimits_{j=1}^{N} \mathbf{W}_{ij} z_j \right\|^2, \end{aligned} \tag{11}$$

[3]Since a graphlet and its post-embedding vector are one-to-one, we do not discriminate them for ease of expression.

where μ is the regularization parameter, K denotes the number of selected graphlets, $\{z_{s_1}, z_{s_2} \cdots, z_{s_K}\}$ contains the selected graphlets, and $\{s_1, s_2, \cdots, s_K\}$ is the set of indices of the selected graphlets. The first term is the cost function to fix the coordinates of the selected graphlets. The second term requires the reconstructed graphlets share the same local structure with the original ones.

Let $\mathbf{Y} = [y_1, y_2, \cdots, y_N]$, $\mathbf{Z} = [z_1, z_2, \cdots, z_N]$, and $\mathbf{\Lambda}$ be an $N \times N$ diagonal matrix whose diagonal entry $\mathbf{\Lambda}_{ii}$ is 1 if $i \in \{s_1, s_2, \cdots, s_K\}$ and 0 otherwise. Then, the above cost function can be reorganized into a matrix form as:

$$\epsilon(\mathbf{Z}) = \mathrm{tr}((\mathbf{Z} - \mathbf{Y})^T \mathbf{\Lambda} (\mathbf{Z} - \mathbf{Y})) + \mu \, \mathrm{tr}(\mathbf{Z}^T \mathbf{U} \mathbf{Z}), \tag{12}$$

where $\mathbf{U} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W})$. To minimize (12), we set the gradient of $\epsilon(\mathbf{Z})$ to zero and obtain:

$$\mathbf{\Lambda}(\mathbf{Z} - \mathbf{Y}) + \mu \mathbf{U} \mathbf{Z} = 0. \tag{13}$$

Thus, the reconstructed graphlets are given by:

$$\mathbf{Z} = (\mu \mathbf{U} + \mathbf{\Lambda})^{-1} \mathbf{\Lambda} \mathbf{Y}. \tag{14}$$

Based on the derived reconstructed graphlets, the reconstruction error is measured by:

$$\begin{aligned} &\varepsilon(y_{s_1}, \cdots, y_{s_K}) = \|\mathbf{Y} - \mathbf{Z}\|_F^2 = \left\| \mathbf{Y} - (\mu \mathbf{U} + \mathbf{\Lambda})^{-1} \mathbf{\Lambda} \mathbf{X} \right\|_F^2 \\ &= \|(\mu \mathbf{U} + \mathbf{\Lambda}) \mu \mathbf{U} \mathbf{X}\|_F^2, \end{aligned} \tag{15}$$

where $\|\cdot\|_F^2$ is the matrix Frobenius norm.

Due to the combinatorial nature, minimizing (15) is computationally heavy on certain computational platforms. To accelerate the learning process, a sequential graphlet discovering scenario is developed. Denote a set of selected graphlets in an image as $\{y_{s_1}, \cdots, y_{s_{K'}}\}$. Let $\mathbf{\Lambda}_n$ be the corresponding $N \times N$ diagonal matrix whose diagonal entry $\mathbf{\Lambda}_{ii} = 1$ if y_i is selected and 0 otherwise, and $\mathbf{\Gamma}_i$ an $N \times N$ matrix whose ii-th entry is 1 and all the others 0. The $s_{K'+1}$-th graphlet is determined by solving:

$$s_{K'+1} = \arg\min_{i \notin \{s_1, \cdots, s_{K'}\}} \left\| (\mu \mathbf{U} + \mathbf{\Lambda}_n + \mathbf{\Gamma}_i)^{-1} \mu \mathbf{U} \mathbf{X} \right\|_F^2. \tag{16}$$

Since matrix $\mathbf{U}$ in (16) is sparse, to accelerate the computation of matrix inversion, based on the Sherman-Morrison-Woodbury formula [5], we obtain:

$$(\mu \mathbf{U} + \mathbf{\Lambda}_n + \mathbf{\Gamma}_i)^{-1} = \mathbf{H} - \frac{\mathbf{H}_{*i} \mathbf{H}_{i*}}{1 + \mathbf{H}_{ii}}, \tag{17}$$

where $\mathbf{H}_{*i}$ and $\mathbf{H}_{i*}$ are the i-th column and the i-th row of $\mathbf{H}$ respectively, and the objective function in (16) can be reorganized as follows:

$$\begin{aligned} &\left\| (\mu \mathbf{U} + \mathbf{\Lambda}_n + \mathbf{\Gamma}_i)^{-1} \mu \mathbf{U} \mathbf{Y} \right\|_F^2 = \mu^2 tr(\mathbf{H} \mathbf{U} \mathbf{Y} \mathbf{Y}^T \mathbf{U} \mathbf{H}) \\ &- \frac{2\mu^2 \mathbf{U} \mathbf{Y} \mathbf{Y}^T \mathbf{U} \mathbf{H} \mathbf{H}_{*i}}{1 + \mathbf{H}_{ii}} + \frac{\mu^2 \mathbf{H}_{i*} \mathbf{H}_{*i} \mathbf{U} \mathbf{Y} \mathbf{Y}^T \mathbf{U} \mathbf{H}_{*i}}{(1 + \mathbf{H}_{ii})^2}. \end{aligned} \tag{18}$$

Denote $\mathbf{A} = \mathbf{U}\mathbf{Y}\mathbf{Y}^T\mathbf{U}$, the optimization problem in (16) can be rewritten as (19) at the bottom of the page.

D. Representative Graphlet Cut

The selected representative graphlets effectively capture the homogeneity among superpixels, which are subsequently incorporated in the normalized cut framework for segmentation. Our proposed approach improves the conventional normalized cut in that: the conventional normalized cut measures the similarity between superpixels using the distance between their appearance feature vectors, whereas our approach measures their similarity by taking into consideration of their spatial structures. Particularly, we train a standard GMM to model their distribution. Given a post-embedding graphlet $f(G_{test})$ from the test image, the homogeneity of its superpixels is computed via:

$$p(f(G_{test})|\theta) = \sum\nolimits_{i=1}^{\mathcal{K}} w_i \mathcal{N}(f(G_{test})|\mu_i, \Sigma_i), \tag{20}$$

where $\theta = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{\mathcal{K}}$ are the GMM parameters learned by using expectation maximization from the training post-embedding graphlets, and we set GMM component number $\mathcal{K} = 5$ in our approach.

On the basis of the above homogeneity measure, the objective function of the proposed graphlet-guided normalized cut is given below:

$$AGcut(V_1, V_2) = \frac{cut(V_1, V_2)}{assoc(V, V_1)} + \frac{cut(V_1, V_2)}{assoc(V, V_2)}, \tag{21}$$

where V_1 and V_2 are two disjoint sets of superpixels. The three terms $cut(V_1, V_2)$, $assoc(V, V_1)$, and $assoc(V, V_2)$ are defined in the following.

The numerator in (21) measures the cost of removing all edges spanning superpixel sets V_1 and V_2, i.e.,

$$\begin{aligned} cut(V_1, V_2) &= \sum\nolimits_{u\in V_1, v\in V_2} w(u, v) \\ &= \frac{1}{|G|} \sum\nolimits_{u\in V_1, v\in V_2} \sum\nolimits_{G\supseteq(u,v)} p(G|\theta), \end{aligned} \tag{22}$$

where $w(u, v)$ is the relationship of superpixel u and v. The term $G \supseteq (u, v)$ collects the parent graphlets of superpixel pair (u, v), and $1/|G|$ functions as a normalization factor.

The two denominators in (21) respectively accumulate connections from superpixels in set V_1 and V_2 to the entire superpixels, i.e.,

$$\begin{aligned} assoc(V, V_1) &= \sum_{u\in V, v\in V_1} w(u, v) \\ &= \frac{1}{|G|} \sum_{u\in V, v\in V_1} \sum_{G\supseteq(u,v)} p(G|\theta), \end{aligned} \tag{23}$$

TABLE I
AVERAGE PER-CLASS MEASURE FROM THE FIVE COMPARED METHODS

	MIM	GMIM	TB	HCRF	Our
SIFT-flow	14%	21%	24%	31.22%	**27.73%**
MSRC-21	67%	73%	77.64%	78.89%	**81.11%**
VOC 2008	8.11%	9.24%	13.22%	20.13 %	**30.12%**
VOC 2009	38.27%	39.16%	41.25%	42.43 %	**43.37%**
VOC 2010	28.43%	29.71%	30.12%	30.13 %	**32.14%**

$$\begin{aligned} assoc(V, V_2) &= \sum_{u\in V, v\in V_2} w(u, v) \\ &= \frac{1}{|G|} \sum_{u\in V, v\in V_2} \sum_{G\supseteq(u,v)} p(G|\theta). \end{aligned} \tag{24}$$

By minimizing the objective function in (21), each test image can be decomposed into several segmented regions. To annotate the semantics of each region, we first learn a multi-label SVM based on the selected d-dimensional post-embedding graphlets and the category labels of the images from which the graphlets are extracted. Given a test graphlet G_{test}, based on the probabilistic output of SVM [16], we obtain its probability of belonging to semantic class c: $p(G_{test} \rightarrow c)$, and the semantic label of segmented region R is computed by maximum majority voting of all its spatially overlapping graphlets:

$$\arg\max_{c} \sum\nolimits_{G_{test}\cap R\neq\emptyset} p(G_{test} \rightarrow c). \tag{25}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section validates the effectiveness of the proposed approach for weakly-supervised segmentation based on four sets of experiments. The first set of experiments compares our approach with representative segmentation algorithms. The second set of experiments evaluates the individual components of our approach. Discussion of parameter setting is given in the third part.

A. Description of Data Sets

To compare our approach with the existing weakly-supervised segmentation methods, we experiment on the SIFT-flow [11] and the MSRC-21 [18]. In addition, it is important to compare our approach with fully-supervised segmentation, as the comparative results show how effectively the image-level labels enhance segmentation performance. To this end, we also experiment on the PASCAL VOC 2008 [4], 2009, and 2010. Note that, only the foreground objects are annotated in images from the VOC series, and we use them as foreground image-level labels. To obtain the background image-level labels, we manually label the background of each image as one of $\{\text{sky}, \text{road}, \text{indoor}\}$, and further combine the foreground image-level labels and the background ones.

$$s_{K'+1} = \arg\min_{i\notin\{s_1,\cdots,s_{K'}\}} \frac{1}{1+\mathbf{H}_{ii}} \left(\frac{\mathbf{H}_{i*}\mathbf{H}_{*i}\mathbf{H}_{i*}\mathbf{A}\mathbf{H}_{*i}}{1+\mathbf{H}_{ii}} - 2\mathbf{H}_{i*}\mathbf{A}\mathbf{H}\mathbf{H}_{*i} \right) \tag{19}$$

TABLE II
AVERAGE PER-CATEGORY ACCURACY ON PASCAL VOC 2008, 2009, AND 2010

Category	aeroplane	bicycle	bird	boat	bottle	bus	car	chair	cow	diningt
VOC 2008	38.14%	39.21%	28.12%	39.47%	32.78%	31.67%	28.11%	31.74%	27.59%	27.35%
VOC 2009	51.37%	43.46%	41.97%	53.49%	46.21%	47.31%	41.69%	45.77%	43.12%	43.21%
VOC 2010	42.21%	31.26%	32.13%	44.43%	35.79%	38.96%	30.97%	34.42%	32.17%	33.39%
Category	diningt.	dog	horse	motorbike	person	pottedp.	sheep	sofa	train	tv
VOC 2008	30.26%	32.78%	24.87%	30.58%	24.43%	30.72%	32.77%	31.53%	23.64%	26.14%
VOC 2009	44.97%	46.75%	39.54%	46.11%	38.87%	44.47%	47.32%	45.59%	38.21%	40.97%
VOC 2010	35.12%	36.36%	29.17%	35.78%	28.64%	35.21%	36.78%	34.78%	27.78%	31.25%

B. Comparison With the State of the Art

In this experiment, we compare our approach with four segmentation methods, including two weakly-supervised segmentation methods: multi-image model (MIM) [22] and its variant (GMIM) [24], as well as two fully-supervised segmentation algorithms: TextonBoost (TB) [18] and hierarchical conditional random field (HCRF) [13].

The experimental settings of the proposed method and its competitors are as follows: For our algorithm, the maximum graphlet size T is set to 5, because we experimentally find that segmentation accuracy improves very limited when $T > 5$. The upper bound of the number of selected graphlets is 1000, making the segmentation carried out quickly. For the four compared methods, MIM, HCRF, and TB are with publicly available codes. For MIM, we use the Matlab toolbox.[4] For HCRF, there are publicly available C++ code.[5] For textonboost, the codes are also publicly downloadable.[6] For all the three algorithms, we kept the parameters in the codes unchanged. For GMIM, we re-implement it based on the publicly available code MIM, because MIM can be deemed as a reduced version GMIM. Toward a fair comparison, the superpxiel generation and the corresponding parameter is the same as the implementation of our algorithm; besides, the features we extracted from each superpixel are also color moment combined with HOG. For the GP optimization, in our implementation, we found that parameter used in the publication [24] is not effective. Actually, we set $mm(\alpha) = 0.3 \sim 0.6$, while the squared exponential is the same as that in the publication.

The segmentation performance is evaluated by average-per-class measure, which averages the correctly classified pixels per-class over all classes (per-category segmentation accuracies of the PASCAL VOC series are given in Table II). In Table I, we report the performance of the five compared methods and two observations are made.

- On both data sets, our approach significantly outperforms the other two weakly-supervised segmentation methods: MIM and GMIM, demonstrating that image-level labels are more effectively encoded by our model.
- Our approach outperforms TextonBoost on both data sets, and performs competitively to HCRF on the PASCAL VOC series. This demonstrates that, even though image-level labels are much coarser cues compared with pixel-level labels, if exploited effectively they can boost segmentation performance to the same extent as pixel-level labels.

[4]http://www.inf.ethz.ch/personal/vezhneva/#code

[5]http://www.inf.ethz.ch/personal/ladickyl/

[6]http://jamie.shotton.org/work/code.html

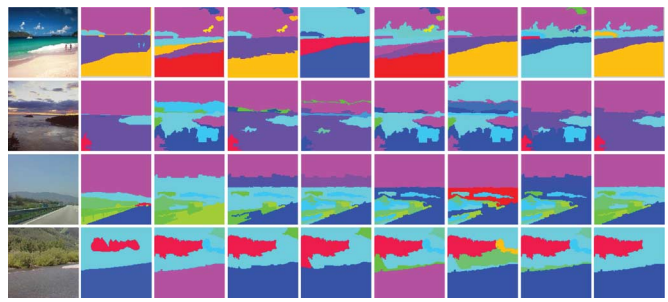

Fig. 5. Example of segmentation results under functionally reduced component (The first column: the original photo, the second column: the ground truth, the third column: superpixel→graphlet, the fourth column: non-structural graphlets, the fifth column: remove image-level labels in the embedding, the sixth column: graphlet embedding→kernel PCA, the seventh column: remove representative graphlet selection, the eighth column: graphlet cut→2-sized graphlets, and the last column: the proposed method).

C. Step-by-Step Model Justification

This experiment evaluates the effectiveness of the four main components in our approach: graphlet extraction, manifold graphlet embedding, representative graphlet selection, and the probabilistic segmentation model.

To justify the effectiveness of graphlets for weakly-supervised segmentation, two experimental settings are adopted to weaken the descriptiveness of graphlets. First, we reduce graphlets to superpixels ("Superpixel as graphlet" in Table III), that is, 1-sized graphlet that captures no spatial structure of superpixels. Second, we remove the structure term $\mathbf{M}_s$ from (3) ("Non-structural graphlet" in Table III). In Fig. 5, we present the segmentation results under the two experimental settings. As shown, segmentation using superpixels or non-structural graphlets results in numerous ambiguous segmentation boundaries.

To justify the effectiveness of manifold graphlet embedding, three experimental settings are used. In the first setting, we remove the geometric context term $\phi_i^h \phi_j^h$ and $\phi_i \phi_j$ from the objective function (5) ("Remove geometric context term" in Table III). In the second setting, we transform our approach into an unsupervised version by abandoning the image-level label encoding term from (5) ("Remove image-level label term" in Table III). In the third setting we transform our approach into an unsupervised version by replacing the manifold embedding with kernel PCA, where the kernel is defined as $k(\mathbf{M}, \mathbf{M}') = \|\mathbf{M}^T \mathbf{M}'\|_F^2$ ("Replace graphlet embedding with kernel PCA" in Table III). We present the segmentation results under the three experimental settings in Fig. 5. By comparing with the ground truth, we can see that removing the geometric

TABLE III
PERFORMANCE DECREASE OF COMPONENT REPLACEMENT

Component replacement	SIFT-flow	MSRC-21	VOC 2008	VOC 2009	VOC 2010
Superpixel as graphlet	4.13%	6.54%	3.76%	5.43%	3.36%
Non-structral graphlet	4.03%	5.78%	3.09%	2.67%	2.71%
Remove geometric context term	6.32%	8.81%	5.31%	6.01%	6.13%
Remove image-level label term	3.63%	4.23%	2.77%	3.54%	3.47%
Replace graphlet embedding with kernel PCA	5.41%	6.56%	4.12%	5.09%	4.16%
Non represe. graphlet selection	7.05%	7.98%	6.14%	6.21%	5.97%
Normalized cut with 2-sized graphlets	4.43%	5.59%	4.21%	4.11%	3.31%

context term results in large number of incorrectly labeled regions. This demonstrates the importance of geometric context into the segmentation process. Besides, segmentation without image-level label supervision performs less satisfactorily, reflecting image-level labels contribute positively to image segmentation. Furthermore, very poor segmentation results are observed when kernel PCA is adopted because both geometric context and image-level labels are abandoned.

To justify the effectiveness of the representative graphlet selection, we abandon this function in the proposed framework and use all the post-embedding graphlets for probabilistic graphlet cut ("Non repres. graphlet selection" in Table III). As seen from Fig. 5, segmentation without the selected graphlets produces several mistakenly annotated regions, which demonstrates the necessity of removing graphlets that are non-representative to a semantic category.

To justify the effectiveness of the probabilistic segmentation model, we restrict the graphlet size to two and thus only binary relationships of superpixels are exploited in the normalized cut based segmentation ("Normalized cut with 2-sized graphlets" in Table III). As shown in Fig. 5, segmentation with 2-sized graphlets results in numerous over-segmented patches, because of the limited superpixel label smoothing capability of 2-sized graphlets. Beyond the analysis of the sample segmentation results, the statistics in Table III shows the performance degradation caused by the above component replacements, again demonstrating the indispensability and inseparability of the four components in our approach.

D. Effects of Different Parameters

The maximum graphlet size T significantly influences the segmentation results. In Fig. 6, we present segmentation accuracy, time consumption corresponding to T ranging from 1 to 10, on the PASCAL VOC 2008 [4]. We do not experiment with T larger than 10 because the segmentation accuracy becomes stable, and the segmentation takes too long, for example, longer than one hour to segment an 1024×768 image. From Fig. 6, we have two observations. First, segmentation accuracy increases moderately as T goes up from 1 to 6, and remains stable as T goes up further from 7 to 10. This implies that 6-sized graphlets are adequately descriptive to capture the homogeneity of superpixels. Second, segmentation time increases exponentially as the graphlet size goes up. Therefore it is better to keep T small.

Next, we present in Fig. 7 the average segmentation accuracy with the dimensionality of post-embedding graphlets ranging from 10 to 130, with a step of 10, again on the SIFT-flow [11]. The maximum dimensionality is set to 130 because the graphlet embedding combines color and texture channel descriptors,

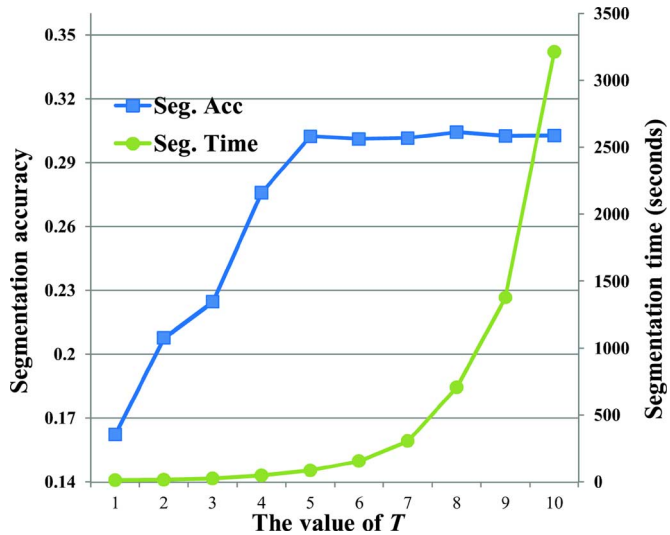

Fig. 6. Segmentation accuracy and time consumption per image under different maximum graphlet size T.

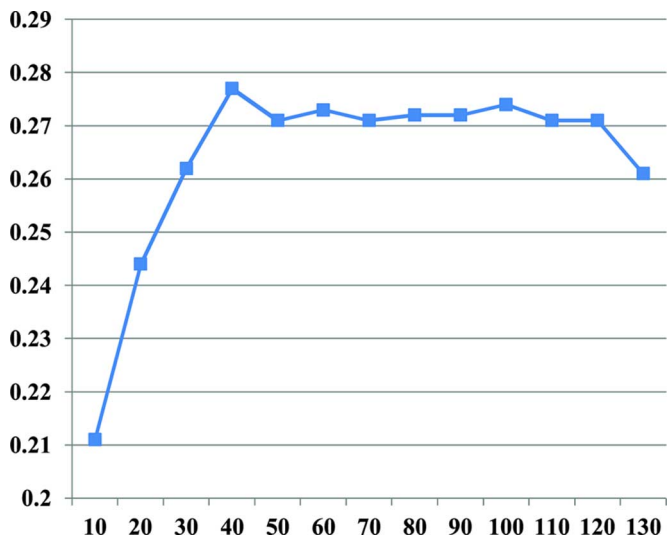

Fig. 7. Average segmentation accuracy with different values of the dimensionality of post-embedding graphlets.

leading to a 137-dimensional vector. As shown in Fig. 7, the best segmentation results are achieved when d is between 40 and 60, and we therefore set $d = 50$ on this data set.

Finally, we present the segmentation performance under different number of those selected graphlets in Fig. 8. Particularly, we set the maximum graphlet size T respectively to $\{4, 5, 6, 7, 8\}$, and obtain $\{211, 432, 851, 1726, 3257\}$ graphlets correspondingly. In each case, we selected

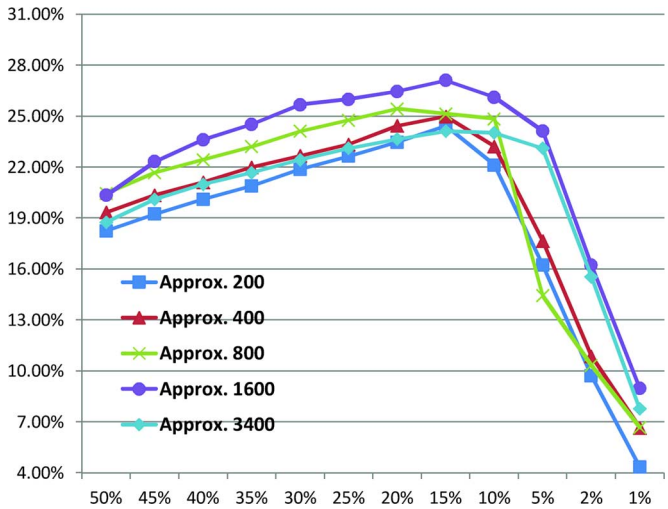

Fig. 8. Segmentation accuracy on SIFT-flow under different number of those selected graphlets.

$\{1\%, 2\%, 5\%, 10\%, 15\%, 20\%, 25\%, 30\%, 35\%, 40\%, 45\%, 50\%\}$ graphlets and reported the segmentation accuracy. As seen, the best segmentation performance is consistently observed when 15% graphlets are selected, which is in line with our opinion that there are numerous graphlets not beneficial to the segmentation task. At the same time, when very few graphlets are selected, the segmentation performance is also sub-optimal. This is because the selected graphlets cannot cover all superpixels.

V. Conclusions and Future Work

This paper presents a weakly-supervised segmentation method by learning the distribution of spatially structured superpixel sets. We introduce the notion of graphlet that captures the spatial structures of superpixels. To integrate image-level labels, a manifold embedding technique is proposed to transform different-sized graphlets into equal-lengthed feature vectors. Based on the embedding, we propose an feature selection algorithm to select a few highly semantics correlated post-embedding graphlets. The selected post-embedding graphlets allow us to use GMM to learn their distribution. The distribution is further used to measure the homogeneity of superpixels for segmenting test images.

In the future, we will investigate a semi-supervised [10] segmentation framework that simultaneously decomposes an image into regions and derives their semantics. In addition, we plan to generalize the proposed method into a multi-level spatial pyramid framework, in order to capture differently-sized objects.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[2] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. ACM Multimedia*, 2010, pp. 291–300.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, pp. 886–893.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computation*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.

[6] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *Proc. CVPR*, 2009, pp. 1–8.

[7] S. Kim, S. Nowozin, P. Kohli, and C. D. Yoo, "Higher-order correlation clustering for image segmentation," in *Proc. NIPS*, 2011, pp. 1530–1538.

[8] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[9] Y. Keselman and S. J. Dickinson, "Generic model abstraction from examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 7, pp. 1141–1156, 2005.

[10] X. Liu, M. Song, D. Tao, Z. Liu, L. Zhang, J. Bu, and C. Chen, "Semi-supervised node splitting for random forest construction," in *Proc. CVPR*, 2013.

[11] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proc. CVPR*, 2009, pp. 1972–1979.

[12] Y. Li, B. Geng, D. Tao, Z.-J. Zha, L. Yang, and C. Xu, "Difficulty guided image retrieval using linear multiple feature embedding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1618–1630, 2012.

[13] L. Ladický, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical CRFs for object class image segmentation," in *Proc. ICCV*, 2009, pp. 739–746.

[14] Z. Ma, F. Nie, Y. Yang, J. Uijlings, N. Sebe, and A. G. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, 2012.

[15] B. Ni, M. Xu, B. Cheng, M. Wang, S. Yan, and Q. Tian, "Learning to photograph: A compositional perspective," *Trans. Multimedia*, accepted for publication.

[16] J. Platt, "Probabilistic outputs for SVMs and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[17] S. Rital, "Hypergraph cuts and unsupervised representation for image segmentation," Fundamenta Informaticae pp. 153–179, 2009.

[18] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. ECCV*, 2006, pp. 1–15.

[19] M. Stricker and M. Orengo, "Similarity of color images," *Storage and Retrieval of Image and Video Databases,* pp. 381–392, 1995.

[20] J. J. Verbeek and B. Triggs, "Region classification with Markov field aspect models," in *Proc. CVPR*, 2007, pp. 1–8.

[21] A. Vezhnevets and J. M. Buhmann, "Towards weakly-supervised semantic segmentation by means of multiple instance and multitask learning," in *Proc. CVPR*, 2010, pp. 3249–3256.

[22] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly-supervised semantic segmentation with a multi-image model," in *Proc. ICCV*, 2011, pp. 643–650.

[23] A. Vezhnevets, J. M. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *Proc. CVPR*, 2012, pp. 3162–3169.

[24] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly-supervised structured output learning for semantic segmentation," in *Proc. CVPR*, 2012, pp. 845–852.

[25] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.

[26] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[27] M. Werman and D. Weinshall, "Similarity and affine invariant distances between 2d point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, pp. 810–814, 1995.

[28] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, 2008.

[29] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua, "Interactive video indexing with statistical active learning," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 17–27, 2012.
[30] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2887–2897, 2013.
[31] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet cut: Exploring spatial structure cue for weakly-supervised image segmentation," in *Proc. CVPR*, 2013.
[32] M. S. Castelhano and M. L. Mack, "Viewing task influences eye movement control during active scene perception," *J. Vision*, vol. 9, no. 3, p. 6, 2009.

Luming Zhang received his Ph.D. degree in computer science from Zhejiang University, China. Currently he is a Postdoctoral Research Fellow at the School of Computing, National University of Singapore. His research interests include multimedia analysis, image enhancement, and pattern recognition.

Yue Gao received the B.S. degree from Harbin Institute of Technology, Harbin, China, in 2005, and the M.E. degree and Ph.D. degree from Tsinghua University, Beijing, China, in 2008 and 2012 respectively. He had been a visiting scholar at Carnegie Mellon University, where he worked with Dr. Alexander Hauptmann from Oct. 2010 to March 2011, a research intern at National University of Singapore and Intel China Research Center, respectively. He is currently a Research Fellow with the School of Computing, National University of Singapore, where he worked with Prof. Tat-Seng Chua. His research interests include large scale multimedia retrieval, 3D object retrieval and recognition, and live social media analysis.

Dr. Gao is a Guest Editor of Multimedia System Journal and Neurocomputing. He has chaired a session Chair of PCM 2012, MMM 2013, ICIMCS2013, and MMM2014. He is a reviewer for IEEE TIP, TMM, CSVT, TIE, and TSMC PART B. He is a member of ACM.

Yingjie Xia received his Ph.D. degree in computer science from Zhejiang University, China. He has been a Postdoc in the Department of Automation, Shanghai JiaoTong University from 2010 to 2012, supervised by Professor Yuncai Liu. Before that, he had been a visiting student at University of Illinois at Urbana-Champaign from 2008 to 2009, supervised by Professor Shaowen Wang. He is currently an associate professor in Hangzhou Institute of Service Engineering, Hangzhou Normal University. His research interests include multimedia analysis, pattern recognition, and intelligent transportation systems.

Ke Lu was born in Guyuan Ningxia on March 13, 1971. He graduated from Department of Mathematics at Ningxia University in July 1993. He received Master degree and Ph.D. degree from Department of Mathematics and Department of Computer Science at Northwest University in July 1998 and July 2003, respectively. He was a Postdoctoral Fellow in Institute of Automation Chinese Academy of Sciences from July 2003 to April 2005. Currently he is a professor of University of the Chinese Academy of Sciences. His research focuses on curve matching, 3D image reconstruction and computer graphics.

Jiale Shen is an Assistant Professor in Information Systems, School of Information Systems, Singapore Management University, Singapore. He received his Ph.D. in computer science from the University of New South Wales (UNSW), Australia in the area of large-scale media retrieval and database access methods. Dr. Shen's main research interests include information retrieval, multimedia systems and economic-aware media analytics. His recent work has been published or is forthcoming in leading journals and international conferences including ACM SIGIR, ACM Multimedia, ACM SIGMOD, ICDE, IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT), IEEE Transactions on Multimedia (IEEE TMM), IEEE Transactions on Image Processing (IEEE TIP), ACM Multimedia Systems Journal, ACM Transactions on Internet Technology (ACM TOIT) and ACM Transactions on Information Systems (ACM TOIS). Besides being chair, PC member, reviewer and guest editor for several leading information systems journals and conferences, he is an associate editor of International Journal of Image and Graphics (IJIG) and area editor of Electronic Commerce Research and Applications (ECRA).

Rongrong Ji received his Ph.D. degree in computer science from Harbin Institute of Technology, China. He has been a Postdoc research fellow in the Department of Electrical Engineering, Columbia University since 2011, working with Professor Shih-Fu Chang. Before that, he had been a visiting student at University of Texas of San Antonio worked with Professor Qi Tian from 2010.3–2010.5, a research assistant at Peking University worked with Professor Wen Gao from 2010.4–2010.11, a research intern at Microsoft Research Asia worked with Dr. Xing Xie, and had led the multimedia retrieval group at Visual Intelligence Lab, Harbin Institute of Technology from 2007–2010. He is the author of over 80 referred journals and conferences in IJCV, TIP, TMM, TOMCCAP, IEEE Multimedia, PR, ACM Multimedia Systems, CVPR, ACM Multimedia, IJCAI, AAAI, etc. His research interests include image and video search, content understanding, mobile visual search, and interactive human-computer interface. Dr. Ji is the recipient of the Best Paper Award at ACM Multimedia 2011 and Microsoft Fellowship 2007. He is an associate editor for International Journal of Computer Applications, a guest editor of International Journal of Advanced Computer Science and Applications, a session chair of ICME 2008, ICIMCS 2010, MMM 2013 and PCM 2012. He serves as reviewer for IEEE TPAMI, TIP, TMM, CSVT, TSMC PART A, B, C, and IEEE Signal Processing Magazine, etc. He is in the program committees of over 20 international conferences including CVPR 2013, ECCV 2012, ACM Multimedia 2012, ACM Multimedia 2011, ICME 2012, etc. He is the guest editor for IEEE Multimedia Magazine.