BOOSTING OBJECTNESS: SEMI-SUPERVISED LEARNING FOR OBJECT DETECTION AND SEGMENTATION IN MULTI-VIEW IMAGES

Huiling Wang[*]

Lappeenranta University of Technology
Finland

Tinghuai Wang

Nokia Technologies
Finland

ABSTRACT

This paper presents a method to detect and segment recurring object from multi-view images. Given a sequence of images of an object captured by multiple cameras, the method firstly detects sparse object-like regions utilizing generic region proposals. We propose a semi-supervised framework to exploit both appearance cues learned from rudimentary detections of object-like regions, and the intrinsic geometric structures within multi-view data. This framework generates a diverse set of object proposals in all views which underpins a robust object segmentation method to handle objects with complex shape and topologies, as well as scenarios where the object and background exhibit similar color distributions.

Index Terms— Multi-view, object detection, segmentation

1. INTRODUCTION

Detecting and segmenting an object captured from multiple viewpoints have gained interest in recent years due to the proliferation of imaging devices. Stable and accurate multi-view object segmentation is fundamental to many computer vision applications, such as 3D reconstruction, image editing, encoding and post-production. This is a very challenging task due to the lack of prior knowledge about object appearance, shape or position. Furthermore, variance in pose, illumination and occlusion relationships introduce ambiguities that in turn induce the potential for localized under- or over-segmentation.

Existing methods for segmenting object in multi-view images are mostly interactive approaches [1, 2, 3, 4]. Yet fully automatic methods remain useful in scenarios where the human in the loop is impractical, such as large scale user-generated content processing. Prior automatic algorithms [5, 6, 7, 8, 9, 10, 11] typically draw upon geometric constraints implicit within multi-view scenario. These methods either simultaneously derive segmentation whilst performing costly visual hull estimation [5], or rely on the color distributions of the object and background being very different [6, 7, 10, 11], or require the fixation of cameras on the object [6, 8], or dense depth recovery of the whole scene [9]. Aside from those associated limitations and challenges, these methods lack an explicit notion of what an object should look like. Consequently, the low-level grouping of pixels usually results in mis-segmentation.

In contrast to previous techniques, our algorithm learns and extracts object proposals from scratch to account for the variation of object's appearance across viewpoints and scene clutters, as opposed to performing low-level grouping of pixels based on color or depth. Our strategy is to create feature-based rudimentary detections of regions for the object by learning from weakly labelled examples of object-like regions. These detections serve as informative indicators of the appearance and location of the object. We propagate this learned prior knowledge on an undirected graph consisting of regions, solving the semi-supervised learning efficiently. Inference at the region level further makes our object proposal extraction approach a practical solution for automatic object segmentation for multi-view images.

2. APPROACH

Our goal is to extract object-like regions from multi-view images, from which we learn appearance and objectness to automatically segment the foreground objects. We achieve our goal in three main steps: (1) discover object-like regions (2) generate a cohort of object proposals by propagating prior knowledge in a semi-supervised learning framework (3) perform object segmentation by learning appearance and objectness from proposals.

2.1. Discovering Object-Like Regions

The goal of this step is to discover an initial set of object-like regions from all views. Throughout the discovery process, we maintain two disjoint sets of image regions: $\mathcal{H}$ and $\mathcal{U}$, which represent the discovered object-like regions and those remain in the general unlabeled pool, respectively. $\mathcal{H}$ is initially empty whilst $\mathcal{U}$ is set to be the regions of all images. Since we assume no prior knowledge on the size, shape, appearance or location of the object, our algorithm operates by producing a diverse set of object-like regions in each image using [12] which is a category independent method to identify object-like regions.

To find the most likely object-like regions among the large set of returned regions, we firstly form a candidate pool $\mathcal{C}$ by taking the top N (N=10) highest-scoring regions from each image. The score of each region consists of two parts: (1) appearance score $\mathcal{A}(r)$ of each region r returned from [12] (2)

[*]This work was performed while Huiling Wang was working as a research intern at Nokia Technologies.

the visibility $\mathcal{V}(r)$ of each region r based on the sparse 3D reconstruction. Specifically, each 3D point from SfM [13, 14] has a number of measures, with each measure representing its visibility, 2D location and photometric properties on the corresponding view. Thus we compute the visibility of each region r by accumulating the number of 3D visibility measures that region r encompasses. Let P_r be the set of 3D points which are visible for region r in one of the views, and n_p be the number of visibilities for each 3D point $p \in P_r$. The visibility of region r can be computed as

$$\mathcal{V}(r) = 1 - \exp(-\frac{\sum_{p \in P_r} n_p}{< \sum_{p \in P} n_p >}),$$

where P represents all the 3D points and $< \sum_{p \in P} n_p >$ is the average visibility of all 3D points. This definition of region visibility takes into account of not only the number of visible 3D points in region r, but also the overall visibility of each 3D points. Our assumption is that one region is more likely to be object-like if it contains more stable feature points whose stability is measured by our visibility measure. The total score is the summation of appearance and visibility of each region.

Then we identify groups of object-like regions that may represent a foreground object by performing spectral clustering [15] in $\mathcal{C}$. To perform clustering, we firstly compute the pairwise affinity matrix between all regions r_i and $r_j \in \mathcal{C}$ as

$$W_{r_i,r_j} = \exp(-\frac{\chi^2(h_a(r_i), h_a(r_j))}{2\beta}), \quad (1)$$

where $h_a(r_i)$ and $h_a(r_j)$ are the color histograms of r_i and r_j respectively, and β is the average distance between all regions. All clusters are ranked based on the average score of its comprising regions. The clusters among the highest ranks correspond to the most object-like regions but there may also be noisy regions, which are added to $\mathcal{H}$.

Each object-like region may correspond to different part of the object from particular image, whereas they collectively describe the object from different viewpoints. We devise a discriminative model to learn the appearance of those most likely object regions. The initial set of object-like regions $\mathcal{H}$ form the set positive instances, while negative instances are randomly sampled outside the bounding box of the positive instances. We use this labeled training set to learn linear SVM classifier for two categories. The classifier provides a confidence of class membership taking the features of a region which combines texture and color features, as input. This classifier is then applied to all the unlabeled regions across all the views. After this classification process, each unlabelled region r_i is assigned with a weight Y_i, i.e. the SVM margin. All weights are normalized between -1 and 1, by the sum of positive and negative margins.

2.2. Generating Multi-View Object Hypotheses

The appearance model from Sec. 2.1 provides an informative yet independent and incoherent prediction on each of the unlabelled regions regardless the inherent spatial and geometric structure revealed by both labeled and unlabeled regions. To generate robust multi-view object hypotheses, we adopt a semi-supervised learning approach, exploiting the intrinsic structure within image data, multi-view geometry and the initial local evidence from the holistic object appearance model. Fig. 1 (b) shows the positive predictions of each region, from SVM predictions and semi-supervised learning respectively. The prediction from SVM exhibits unappealing incoherence, nonetheless, using it as initial input, semi-supervised learning gives smooth predictions exploiting the inherent structure of data, as shown in Fig. 1 (c).

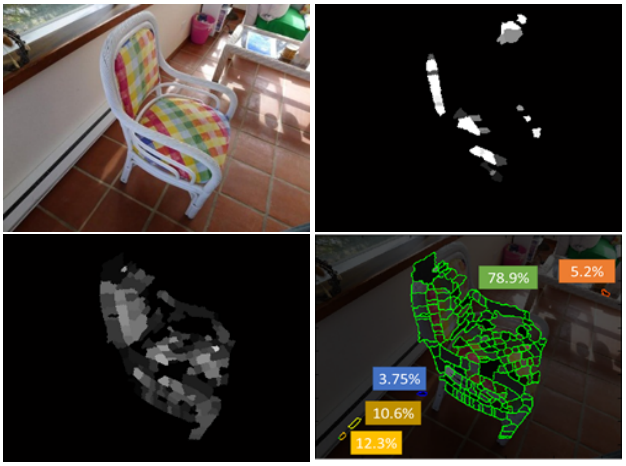

Fig. 1. Multi-View object hypotheses generation (a) source image (b) positive predictions from SVM (c) predictions from semi-supervised learning captures the coherent intrinsic structure within visual data, using SVM predictions as input (d) generated object hypotheses with average objectness values indicated by the brightness.

To perform semi-supervised learning, we define a weighted graph $\mathcal{G}_s = (\mathcal{V}, \mathcal{E})$ spanning all the views with each node corresponding to a region, and each edge connecting two regions based on intra-view and inter-view adjacencies. Intra-view adjacency is defined as the spatial adjacency of regions in the same view whilst inter-view adjacency coarsely determined based on the visibility of reconstructed sparse 3D points from sparse reconstruction. Specifically, the regions which contain 2D projections (2D feature points) of the same 3D point are adjacent. See Fig. 2 for an illustrative description. Note that accurate camera calibration is neither assumed nor required to construct this graph.

We compute the affinity matrix W of the graph using the same formulation in Eq. 1. Since sparsity is important to remove label noise and semi-supervised learning algorithms are more robust on sparse graphs [16], we set all W_{ij} are set to zero if r_i and r_j are not adjacent. Semi-supervised learning propagates label information from labeled nodes to unlabeled nodes. Let the node degree matrix $D = \text{diag}([d_1, \ldots, d_N])$ be defined as $D_i = \sum_{j=1}^{N} W_{ij}$, where $N = |\mathcal{V}|$. We formulate the problem as minimizing an energy function $E(X)$ with respect to all region labels X:

$$E(F) = \sum_{i,j=1}^{N} W_{ij} |\frac{F_i}{\sqrt{D_i}} - \frac{F_j}{\sqrt{D_j}}|^2 + \mu \sum_{i=1}^{N} |F_i - Y_i|^2, \quad (2)$$

where $\mu > 0$ is the regularization parameter, and Y are the desirable labels of nodes which are normally imposed by prior knowledge. The first term in (2) is the *smoothness constraint*, which encourages the coherence of labelling among adjacent nodes, whilst the second term is the *fitting constraint* which enforces the labelling to be similar with the initial label assignment. The optimization problem in Eq. (2) can be solved by an iteration algorithm in [17, 18, 19]. More efficiently, we solve it as a linear system of equations. Differentiating $E(F)$ with respect to F we have

$$\nabla E(F)|_{F=F^*} = F^* - SF^* + \mu(F^* - Y) = 0 \tag{3}$$

where $S = D^{-1/2}WD^{-1/2}$. It can be transformed as

$$F^* - \frac{1}{1+\mu}SF^* - \frac{\mu}{1+\mu}Y = 0 \tag{4}$$

Denoting $\gamma = \frac{\mu}{1+\mu}$, we have $(I - (1-\gamma)S)F^* = \gamma Y$. An optimal solution for F can be solved using the Conjugate Gradient method with very fast convergence. We use the predictions from SVM classifier to assign the values of Y. The diffusion process can be performed for positive and negative labels separately, with initial labels Y in Eq. 2 substituted as Y_+ and Y_- respectively:

$$Y_+ = \begin{cases} Y & \text{if } Y > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and

$$Y_- = \begin{cases} -Y & \text{if } Y < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Combining the diffusion processes of both the object-like regions and background can produce more efficient and coherent labelling, taking advantage of their complementary properties. We perform the optimization for two diffusion processes simultaneously as follows:

$$F^* = \gamma(I - (1-\gamma)S)^{-1}(Y_+ - Y_-). \tag{7}$$

This enables a faster and stable optimization avoiding separate optimizations while giving equivalent results to the individual positive and negative label diffusion.

Finally, the regions which are assigned with label $F > 0$ from each view are grouped. Specifically, we use the final label F to indicate the level of objectness of each region. The final proposals are generated by grouping the spatially adjacent regions ($F > 0$), and assigned by an objectness value by averaging the constituent region-wise objectness F weighted by area. The grouped regions with the highest objectness per view are added to the set of object proposals $\mathcal{P}$. Exemplar object proposals are shown in Fig. 1 (d).

2.3. Multi-View Object Segmentation

We formulate multiple view segmentation as a pixel-labelling problem of assigning each pixel with a binary value which represents background or object respectively. We define a graph by connecting pixels spatially corresponding to the same 3D sparse points, which is similar to the region-based graph case in Sec. 2.2. In contrast to the previous graph during semi-supervised learning, each of the nodes in this graph is a pixel as opposed to a region. We define the energy function that minimizes to achieve the optimal labeling:

$$E(x) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \lambda \sum_{i \in \mathcal{V}, j \in N_i} \psi_{i,j}(x_i, x_j) \tag{8}$$

where N_i is the set of pixels adjacent to pixel i in the graph and λ is a parameter.

The pairwise term $\psi_{i,j}(x_i, x_j)$ penalizes different labels assigned to adjacent pixels:

$$\psi_{i,j}(x_i, x_j) = [x_i \neq x_j]\text{exp}(-d(x_i, x_j))$$

where $[\cdot]$ denotes the indicator function. The function $d(x_i, x_j)$ computes the color and edge distance between neighboring pixels:

$$d(x_i, x_j) = \beta(1 + |SE(x_i) - SE(x_j)|) \cdot ||c_i - c_j||^2$$

where $SE(x_i)$ ($SE(x_i) \in [0, 1]$) returns the edge probability provided by the Structured Edge (SE) detector [20], $||c_i - c_j||^2$ is the squared Euclidean distance between two adjacent pixels in CIE Lab colorspace, and $\beta = (2 < ||c_i - c_j||^2 >)^{-1}$ with $< \cdot >$ denoting the expectation.

The unary term $\psi_i(x_i)$ defines the cost of assigning label $x_i \in \{0, 1\}$ to pixel i, which is defined based on the per-pixel probability map by combining color distribution and region objectness:

$$\psi_i(x_i) = -\text{log}(w \cdot U_i^c(x_i) + (1 - w) \cdot U_i^o(x_i))$$

where $U_i^c(\cdot)$ is the color likelihood and $U_i^o(\cdot)$ is the objectness cue. We adopt the binary graph cut [21] to minimize Eq. 8 and the resulting label assignment gives the foreground object segmentations of the multi-view images.

We estimate two Gaussian Mixture Models (GMM) in CIE Lab colorspace to model the appearance of the object and background. Pixels belonging to the set of object proposals are used to train the GMM of the object, whilst randomly sampled pixels in the complement of object proposals are adopted to train the GMM for the background. Given the estimated GMM color models, per-pixel probability $U_i^c(\cdot)$ is defined as the likelihood observing each pixel as object or background respectively can be computed.

The extracted object proposals provide explicit information of how likely a region belongs to the foreground object (objectness) which can be directly used to drive the final segmentation. We set the per-pixel likelihood $U_i^o(\cdot)$ to be related to the objectness value (F in Eq. 2) of the region it belongs to:

$$U_i^o(x_i) = \begin{cases} F_i & \text{if } x_i = 1 \\ 1 - F_i & \text{if } x_i = 0 \end{cases} \tag{9}$$

3. RESULTS

For implementation, we empirically set $\mu = 3.0$ to balance the impact of the prior labelling and the local labelling smoothness. For graph cut optimization, we empirically set $\lambda = 5$ and $w = 0.35$. These parameters are fixed for the evaluation.

Table 1. Quantitative results on four datasets. The proposed method is compared with four state-of-the-art methods.

Dataset	Our method	Djelouah [11]	Kowdle [9]	Djelouah [10]	Vicente [22]
COUCH	99.7±0.2	99.0±0.2	99.6±0.1	98.8±0.8	NA
TEDDY	98.6±0.3	98.0±1.0	98.8±0.4	98.8±0.4	NA
CHAIR1	99.4±0.2	98.6±0.3	99.2±0.4	88.0±0.2	86.9±7.8
CAR	98.5±0.5	97.0±0.8	98.0±0.7	NA	91.4±4.3

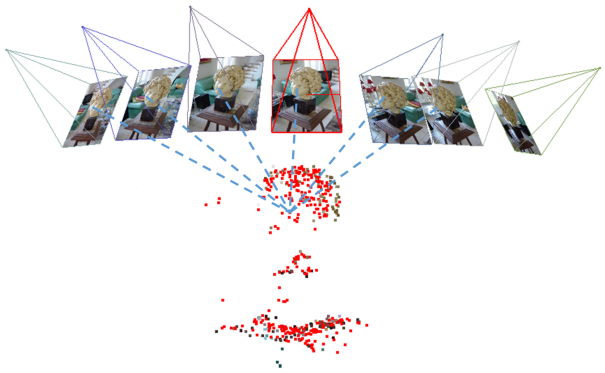

Fig. 2. Sparse 3D reconstruction and rough camera pose using Structure from Motion (SfM). Regions or pixels in views containing the 2D projection of the same 3D point are deemed adjacent in the graph.

To demonstrate the efficacy of our approach, we perform evaluations on challenging datasets. Note that the scarcity of publicly available multi-view image datasets made the comparisons difficult. We obtained five datasets from two state-of-the-art methods: BUSTE dataset from [7] for qualitative evaluation; COUCH, TEDDY, CHAIR1, CAR from [9] which we use for both qualitative and quantitative evaluation. We use the same evaluation metric as [9, 11], computing the intersection over union to measure the segmentation quality.

Fig. 3 shows the qualitative results applying our method on five datasets. The proposed method demonstrates superior segmentation accuracy in challenging scenarios, e.g. in the presence of large illumination variation (row 1), overlapping color distribution and hairy boundary (row 3), complex topology (row 4), and diffused inter-reflections between object and backgrounds (row 5).

As shown in Table 1, our method quantitatively outperforms the competing methods in 3 out of 4 datasets, with relatively less variations on all datasets, which confirms what we observed from qualitative evaluation. Our method substantially outperforms the competing methods on CHAIR1 and CAR which are more challenging than the rest. This indicates that, with an explicit notion of object, our method is more robust in dealing with complex topology and cluttered scene than the traditional low-level pixel group approaches. The stereo based method [9] slightly outperforms our method on TEDDY, owing to the low level pixel grouping as well as depth plane detection which requires smaller camera baseline for the purpose of obtaining the stereo. The method [10] also marginally outperforms our method on TEDDY, possibly due to the low level pixel grouping and accurate camera parameters. Note that accurate camera calibration is neither assumed nor required in our proposed method.

Fig. 3. Qualitative evaluation results on BUSTE (row 1), COUCH (row 2), TEDDY (row 3), CHAIR1 (row 4), and CAR (row 5) datasets.

4. CONCLUSION

We presented an algorithm that automatically discovers recurring object-like regions and generates object hypotheses through a novel semi-supervised learning framework in multi-view images, which in turn underpins a robust object segmentation method. By harnessing a top-down explicit notion of object, our method overcomes the limitations of previous bottom-up methods that often mis-segment an object and delivers high quality segmentation.

5. REFERENCES

[1] Mario Sormann, Christopher Zach, and Konrad F. Karner, "Graph cut based multiple view segmentation for 3d

reconstruction," in *3DPVT*, 2006, pp. 1085–1092.

[2] Jianxiong Xiao, Jingdong Wang, Ping Tan, and Long Quan, "Joint affinity propagation for multiple view segmentation," in *ICCV*, 2007, pp. 1–7.

[3] Adarsh Kowdle, Dhruv Batra, Wen-Chao Chen, and Tsuhan Chen, "imodel: Interactive co-segmentation for object of interest 3d modeling," in *Trends and Topics in Computer Vision - ECCV*, 2010, pp. 211–224.

[4] Tinghuai Wang, John P. Collomosse, and Adrian Hilton, "Wide baseline multi-view video matting using a hybrid markov random field," in *ICPR*, 2014, pp. 136–141.

[5] Gang Zeng and Long Quan, "Silhouette extraction from multiple images of an unknown background," in *ACCV*, 2004.

[6] Neill D. F. Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla, "Automatic 3d object segmentation in multiple views using volumetric graph-cuts," *Image Vision Comput.*, vol. 28, no. 1, pp. 14–25, 2010.

[7] Wonwoo Lee, Woontack Woo, and Edmond Boyer, "Silhouette segmentation in multiple views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1429–1441, 2011.

[8] Neill DF Campbell, George Vogiatzis, Carlos Hernandez, and Roberto Cipolla, "Automatic object segmentation from calibrated images," in *CVMP*. IEEE, 2011, pp. 126–137.

[9] Adarsh Kowdle, Sudipta N. Sinha, and Richard Szeliski, "Multiple view object cosegmentation using appearance and stereo cues," in *ECCV*, 2012, pp. 789–803.

[10] Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Le Clerc, and Patrick Pérez, "N-tuple color segmentation for multi-view silhouette extraction," in *ECCV*, 2012, pp. 818–831.

[11] Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, François Le Clerc, and Patrick Pérez, "Multi-view object segmentation in space and time," in *ICCV*, 2013, pp. 2640–2647.

[12] Ian Endres and Derek Hoiem, "Category independent object proposals," in *ECCV*, 2010, pp. 575–588.

[13] Andrew Harltey and Andrew Zisserman, *Multiple view geometry in computer vision (2. ed.)*, Cambridge University Press, 2006.

[14] Noah Snavely, Steven M. Seitz, and Richard Szeliski, "Photo tourism: exploring photo collections in 3d," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[15] Edwin Olson, Matthew Walter, John Leonard, and Seth Teller, "Single cluster graph partitioning for robotics applications," in *Proceedings of Robotics Science and Systems*, 2005, pp. 265–272.

[16] Tony Jebara, Jun Wang, and Shih-Fu Chang, "Graph construction and *b*-matching for semi-supervised learning," in *ICML*, 2009, p. 56.

[17] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Sch, "Learning with local and global consistency," in *NIPS*, 2004, vol. 1.

[18] Tinghuai Wang and Huiling Wang, "Graph transduction learning of object proposals for video object segmentation," in *ACCV*, pp. 553–568. Springer, 2014.

[19] Huiling Wang and Tinghuai Wang, "Primary object discovery and segmentation in videos via graph-based transductive inference," *Computer Vision and Image Understanding*, 2016.

[20] Piotr Dollar and C. Lawrence Zitnick, "Structured forests for fast edge detection," in *ICCV*, December 2013.

[21] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[22] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov, "Object cosegmentation," in *CVPR*, 2011, pp. 2217–2224.