Joint Semantic Segmentation by Searching for Compatible-Competitive References

Ping Luo[1,3], Xiaogang Wang[2,3], Liang Lin[4]*, and Xiaoou Tang[1,3]†
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Department of Electronic Engineering, The Chinese University of Hong Kong
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[4]Sun Yat-Sen University, Guangzhou, P.R.China
pluo.lhi@gmail.com, xgwang@ee.cuhk.edu.hk, linliang@ieee.org, xtang@ie.cuhk.edu.hk

ABSTRACT

This paper presents a framework for semantically segmenting a target image without tags by searching for references in an image database, where all the images are unsegmented but annotated with tags. We jointly segment the target image and its references by optimizing both *semantic consistencies* within individual images and *correspondences* between the target image and each of its references. In our framework, we first retrieve two types of references with a semantic-driven scheme: i) the compatible references which share similar global appearance with the target image; and ii) the competitive references which have distinct appearance to the target image but similar tags with one of the compatible references. The two types of references have complementary information for assisting the segmentation of the target image. Then we construct a novel graphical representation, in which the vertices are superpixels extracted from the target image and its references. The segmentation problem is posed as labeling all the vertices with the semantic tags obtained from the references. The method is able to label images without the pixel-level annotation and classifier training, and it outperforms the state-of-the-arts approaches on the MSRC-21 database.

Categories and Subject Descriptors

I.4.8 [**Computing Methodologies**]: Image Processing and Computer Vision—*Scene Analysis*

Keywords

semantic segmentation, scene understanding, image search, label propagation

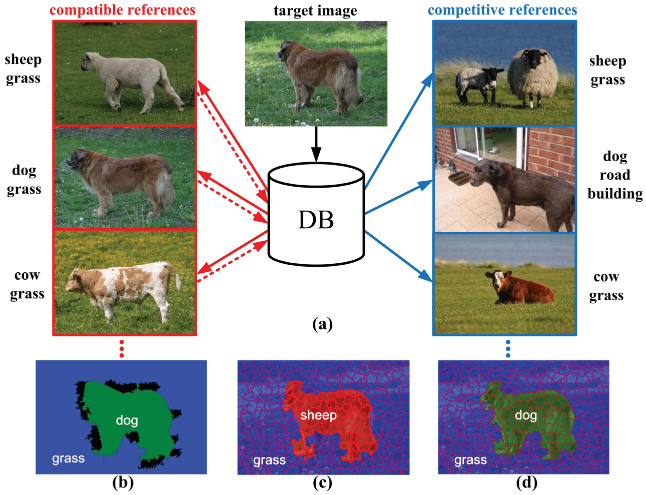

Figure 1: Given a target image without tags, our framework retrieves several compatible-competitive images with tags as references and jointly segments all these images. In (a), each compatible reference has similar global appearances with the target image, while the competitive references vary in appearance but share similar semantic concepts with the compatible references. (b) shows the ground truth label map of the target image. (c) and (d) are the joint segmentation results of using only the compatible references and using both of them respectively. *This figure is encouraged to be view in electronic version.*

1. INTRODUCTION

With the massive explosion of tagged images on the Internet (*e.g.*, Flickr, Google), an interesting question arises recently: How to segment them together with untagged images and simultaneously assign semantic tags to the regions of all the images? This problem is very challenging due to the fact that no pixel-level annotation is available to train a classifier. A few works have been proposed to address this problem [10, 8, 9], and they mainly rely on an assumption that images with similar global appearance tend to share the same semantic concepts. This assumption is not always true due to object/scene variances and ambiguities. As an example shown in Fig. 1, the found reference images with similar global appearance as the target image have different sets of tags and may lead to segmentation errors.

We investigate a novel approach in this paper to make use of conflicting (competitive) relations based on the inconsistency of appearances and semantics. Our idea is illustrated in Fig.1. Given an

[*] This work is partially supported by the National Natural Science Foundation of China (Grant No.61173082).

[†] This work is partially supported by Hong Kong SAR through RGC project 416510, and by Guangdong Province through Introduced Innovative R&D Team of Guangdong Province 201001D0104648280, and Shenzhen Key Lab of Computer Vision and Pattern Recognition.

MM'12, October 29–November 2, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...$15.00.

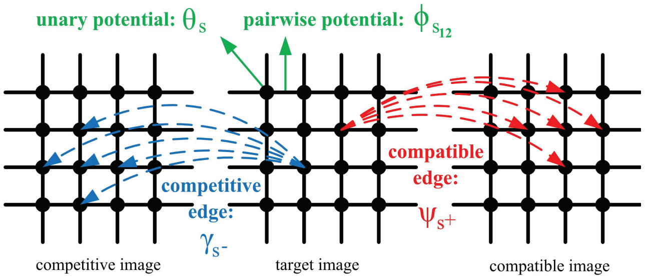
unary potential: θ_s
pairwise potential: $\phi_{s_{12}}$
compatible edge: ψ_{s^+}
competitive edge: γ_{s^-}
competitive image
target image
compatible image

Figure 2: Illustration of cc-MIG. All the superpixels of the target image as well as those of its references are extracted as graph vertices, of which the class labels are inferred. The semantic consistency of each image is represented by the unary and pairwise potentials, *i.e.*the statistical image priors. The image correspondences between the target image and each compatible (or competitive) image are represented by the compatible edges (or competitive edges).

untagged target image shown in the middle, we search for both the **compatible** and **competitive** references, and jointly segment them with the target image to achieve superior performance. Specifically, the compatible references share similar global appearances with the target image, while the competitive references have distinct appearance to the target image but similar tags with one of the compatible references. The correspondences between the superpixels in the target image and in the reference images are established considering the locations and appearance of superpixels. If a superpixel in the target image has similar appearance with its corresponding superpixel in a compatible reference image, they are encouraged to have to same label. This introduces errors if the compatible image and the target image actually have different semantic concepts. Therefore, we further pose a constraint that if a superpixel in the target image has different appearance with its corresponding superpixel in a competitive reference image, they are encouraged to have different labels.

Our framework contains the off-line and on-line phases. During the off-line phase, each image in the database is over-segmented into superpixels, for each of which we compute a 119-dimensional feature vector as described in [4]. Several image priors are then adopted over these superpixels (sec.2.2). During the on-line phase, given an target image, we first search for its compatible references using GIST and HOG. Each of the compatible image is then associated with a competitive reference by a semantic-driven retrieval scheme (sec.2.3). Furthermore, we propose a compatible/competitive multi-image graph (cc-MIG), with which the class label of each superpixel is inferred, as Fig.2 illustrates. The cc-MIG is a combinatorial representation constructed with the target image as well as its references. In this representation, each image creates a single graph, where the unary and pairwise potentials are modeled by the statistical image priors; the correspondences between superpixels in the target image and those in each compatible (or competitive) image are modeled by the **compatible edges** (or **competitive edges**). Each compatible edge, connecting two superpixels from two respective graphs, encourages the superpixels to have the same label, supposing they have similar appearances and spatial locations. By contrast, each competitive edge indicates two connected superpixels with different appearances yet similar locations to be assigned with distinct labels. With the cc-MIG representation, the task of class label assignment can be efficiently solved via a linear programming relaxation (sec.2.1).

Related Work. The jointly semantic segmentation in literature can be divided into two categories: (i) tag propagation and (ii) pixelwise label propagation. The former transfers the tags of training data into unseen images. It requires less labeling effort but is limited by their accuracy. The latter makes use of the full supervision (*i.e.*, pixelwise label map) [5] and achieves the state-of-the-arts performance. Its disadvantage lies in expensive annotations and expansibility.

2. OUR FRAMEWORK

We first present our formulation for the cc-MIG, where the optimizing semantic consistencies and image correspondences of superpixels can be solved as a linear programming problem. Then, we introduce the image priors to model the potential functions of cc-MIG. The semantic-driven scheme of searching references is discussed at the end.

2.1 A Linear Programming Formulation

The objective function can be formulated as summing the scores of semantic consistencies and image correspondences together

$$\max \sum_{\forall I\in I^t\cup\{I_i^+,I_i^-\}} sem(I) + \sum_{\forall<I^{+/-},I^t>} corr(I^{+/-},I^t), \quad (1)$$

where I^t, I^+, I^- denote the target image, compatible reference, and the competitive reference respectively.

Semantic consistency. For all $(s, l_s) \in I$, where s and l_s indicate a superpixel and its label respectively. In our method, we constrain that each superpixel can be assigned only one label. We introduce a binary indicator $x_s \in \mathbb{R}^{|L_s|}$ for each superpixel, where $\sum_{i=1}^{|L_s|} x_s(i) = 1$ and $x_s(i) = 1$ if $i = l_s$. L_s is the set of possible labels of s[1]. Also, for every two connected superpixels s_1 and s_2 in image I, we present a binary indicator $y_{s_{12}} \in \mathbb{R}^{|L_{s_1}|\times|L_{s_2}|}$, which is a matrix and can be defined as $\sum_{i=1}^{|L_{s_1}|}\sum_{j=1}^{|L_{s_2}|} y_{s_{12}}(i,j) = 1$, and $y_{s_{12}}(i,j) = 1$ if $i = l_{s_1}$ and $j = l_{s_2}$.

$sem(I)$ can be then defined as

$$\sum_{\forall s,i=1}^{|L_s|} \theta_s(i)x_s(i) + \sum_{\forall<s_1,s_2>,i,j}^{L_{s_1}\times L_{s_2}} \phi_{s_{12}}(i,j)y_{s_{12}}(i,j), \quad (2)$$

where $\theta(\cdot)$ and $\phi(\cdot,\cdot)$ are the unary and pairwise potentials respectively, which are modeled by statistical image priors and will be discussed in sec.2.2. Note that Eq.(2) is a relaxation of Markov Random Field (MRF). Beside the original constraints for x_s and $y_{s_{12}}$, we also ensure the solution stability among them as

$$\sum_{i}^{|L_{s_1}|} y_{s_{12}}(i,j) = x_{s_2}(j), \sum_{j}^{|L_{s_2}|} y_{s_{12}}(i,j) = x_{s_1}(i). \quad (3)$$

Therefore, the first summation of Eq.(1) along with its constraints can be expressed in the following matrix notation

$$\begin{aligned} &\Theta^T\mathbf{x} + \Phi^T\mathbf{y} \\ s.t. \quad &H\mathbf{x} = \mathbf{e}, A\mathbf{x} = B\mathbf{y},\ and\ x, y \in \{0,1\}, \end{aligned} \quad (4)$$

where $\mathbf{x}$ is a long vector, containing all binary vectors of all superpixels from all the images. In the same manner, we concatenate all binary matrices into $\mathbf{y}$. Moreover, $\mathbf{e}$ is an all one vector, and H, A, B are the coefficient matrices. In fact, Eq.(4) couples all the MRF relaxations.

[1] L_s is determined based on apriori knowledge. For example, for the superpixels of a reference, the possible labels are the given tags.

Image correspondence. Correspondences between images are modeled by compatible and competitive edges in the cc-MIG. For each compatible edge that connects two superpixels s^+ and s[2], we introduce a binary indicator z^+, which is also a matrix and is defined to be $\sum_{i=1}^{|L_{s^+}|}\sum_{j=1}^{|L_s|} z^+(i,j) = 1$, and $z^+(i,j) = 1$ if $i = l_{s^+}$ and $j = l_s$. Moreover, a binary matrix z^- can be established in the same way for each competitive edge. The second summation of Eq.(1) can be decomposed as

$$\sum_{\forall(I^+,I^t)} \sum_{\forall(s^+,s),(i,j)}^{L_{s^+}\times L_s} \psi_{s^+}(i,j)z^+(i,j) + \tag{5}$$

$$\sum_{\forall(I^-,I^t)} \sum_{\forall(s^-,s),(i,j)}^{L_{s^-}\times L_s} \gamma_{s^-}(i,j)z^-(i,j),$$

Here, $\psi_{s^+}(i,j) = exp\{- \parallel f_{s^+} - f_s \parallel_2\}$ if $i = j$ and $\psi_{s^+}(i,j) = 0$ otherwise, and $\gamma_{s^-}(i,j) = exp\{- \parallel f_{s^-} - f_s \parallel_2\}$ if $i \neq j$ and $\gamma_{s^-}(i,j) = 0$ otherwise. f denotes the feature of the superpixel. Imposing Eq.(5) with the constraints similar to Eq.(3), we can also formulate image correspondences in matrix form

$$\Psi^T\mathbf{z}^+ + \Gamma^T\mathbf{z}^- \tag{6}$$

$$s.t. \quad C\mathbf{z}^+ = D\mathbf{x}, C'\mathbf{z}^- = D'\mathbf{x},\ and\, x, z^+, z^- \in \{0,1\},$$

where $\mathbf{z}^+, \mathbf{z}^-$ denote two long vectors by concatenating all binary matrices, and C, C', D, D' are the coefficient matrices. The constraints of Eq.(6) are to guarantee solution stability between variables $\mathbf{z}^+, \mathbf{z}^-$ and $\mathbf{x}$ respectively.

Combining Eq.(4) and Eq.(6) leads to the complete formulation for cc-MIG:

$$\max_{\mathbf{x},\mathbf{y},\mathbf{z}^+,\mathbf{z}^-} \quad \Theta^T\mathbf{x} + \Phi^T\mathbf{y} + \Psi^T\mathbf{z}^+ + \Gamma^T\mathbf{z}^- \tag{7}$$

$$s.t. \quad H\mathbf{x} = \mathbf{e}, A\mathbf{x} = B\mathbf{y}, C\mathbf{z}^+ = D\mathbf{x}, C'\mathbf{z}^- = D'\mathbf{x}$$

$$and \quad x, y, z^+, z^- \in \{0,1\},$$

If we let $0 \leq x \leq 1, x \in \mathbf{x}, \mathbf{y}, \mathbf{z}^+, \mathbf{z}^-$, this integer program is relaxed to a linear one, which is sparse and can be efficiently solved using interior point method.

2.2 Image Priors

Now we describe the potential functions in Eq.(4). The pairwise potential $\phi_{s_{12}}(\cdot,\cdot)$ is modeled with the first-order density prior introduced in [6]. The unary potential $\theta_s(\cdot)$ is modeled with the linear combination of three statistical image priors: *semantics-based superpixel density prior*, *objectness*, and *saliency*. We will explain the first one, which is proposed by us. The second and the third are discussed in [1] and [2] respectively.

Given a superpixel from the input image, we estimate its class label distribution by the semantics-based superpixel density prior, which is found on the basic observation that if a superpixel possess high density in an image, it should probably be assigned one of the tags of this image. Therefore, this prior is computed in the following three steps. i) For each superpixel s from the input image, we first estimate its density on every image I in the database by

$$d_s^I \propto \frac{1}{k}\sum_i^k \exp\{- \parallel f_s - f_{s_i} \parallel_2\}, \forall s_i \in I. \tag{8}$$

Eq.(8) considers the density as the average similarities between superpixel s and its k-nearest neighbors in image I. ii) We then rank the images in a descent order according to their densities of s. iii) At last, we select the first $1/20$ of the total images to calculate the class label distribution of superpixel s.

We verify the above prior on the MSRC-21 dataset, which is split into training and testing sets in the standard way. The class distribution of each testing superpixel is estimated from the training data. We let $k = 50$ and achieve overall 57% classification accuracy given 21 classes in total.

2.3 Semantic Search

We propose a semantic-driven scheme to retrieve the compatible and competitive references from the database. Each compatible reference has small distance in global appearance space to the target image, while the competitive references vary in appearances but share similar semantic concepts with the compatible ones. We first introduce two distance metrics, then discuss our semantic-driven scheme.

The distance metrics for global appearances and semantic similarities are defined as

$$dist^{app}(I^+, I^-) = \parallel f_{I^+} - f_{I^-} \parallel_2 \tag{9a}$$

$$dist^{sem}(I^+, I^-) = 1 - \frac{|T(I^+) \cap T(I^-)|}{|T(I^+) \cup T(I^-)|}, \tag{9b}$$

where f is the feature of combining HOG and GIST, and $T(\cdot)$ denotes the set of the image's tags. Our scheme has two steps. First, using the target image as query, we search its nearest neighbors according to Eq.(9a). The nearest neighbors are treated as compatible references. Second, for each compatible image, we retrieve $1/10$ of the total images in the database in a descent order based on Eq.(9a). Therewith, we apply greedy search to find a competitive image with the smallest value of Eq.(9b). Note that this process is fast due to all the distances can be computed off-line in the database.

3. EXPERIMENTS

Data. We evaluate our method on the MSRC-21 dataset [7], including 591 images with ground truth label maps of 21 classes.

I. Analysis of competitiveness. We validate the effectiveness of utilizing the competitive references and use the default split of training and testing data defined in [7]. The results are shown in Fig.3, where the $x - axis$ indicates how many reference pairs (as explained in sec.2.3, one compatible image is associated with one competitive image) are retrieved, and the $y - axis$ shows the average accuracy of semantic segmentation on the test images. The blue (left) and red (right) bars are the results of only using compatible references and using both types of references.

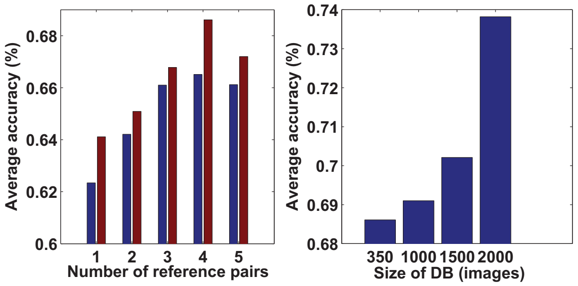

Figure 3: Segmentation accuracies of the experiment I and III.

II. Comparisons. We compare the proposed method with the following state-of-the-art algorithms: MIM [10], PLSA-MRF [8],

[2] We don't attempt to connect all the superpixels between images, since this would produce an over-complex graph. Inspired by [10], only the top k similar/dissimilar superpixel pairs are connected by compatible/competitive edges. $k = 5$ in our experiments.

	average	building	grass	tree	cow	sheep	sky	airplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat
MTL-RF	37	7	96	18	32	6	99	0	46	97	54	74	54	14	9	82	1	28	47	5	0	0
PLSA-MRF	50	45	64	71	75	74	86	81	47	1	73	55	**88**	6	6	63	18	80	27	26	55	8
MIM	67	12	83	70	**81**	93	84	91	55	97	87	92	82	69	51	61	59	66	53	44	9	58
Ours	69	6	80	**72**	**81**	**94**	82	82	**77**	90	85	89	**88**	**72**	**60**	65	**63**	73	**59**	**51**	32	40

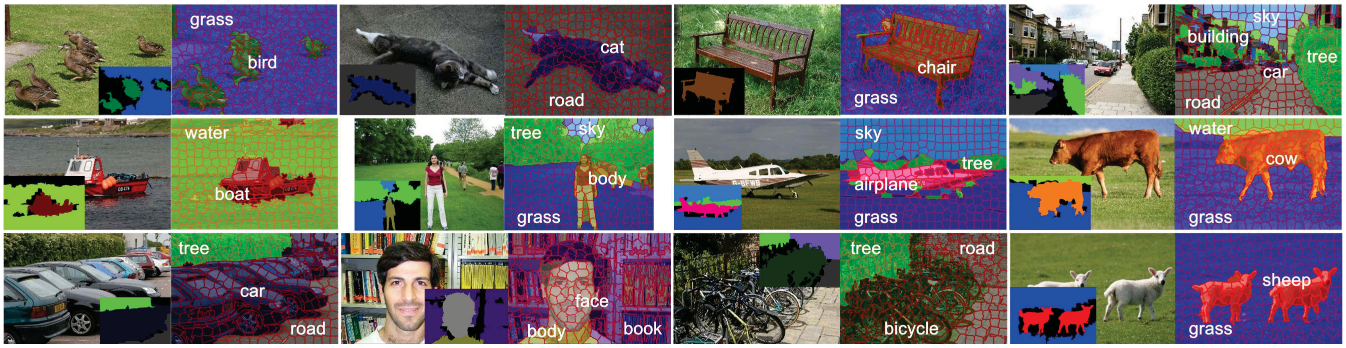

Figure 4: Top: we summarize the results of experiment II compared with the state-of-the-arts. Classes on which we perform better than other methods are shown in bold. Note that our approach works well on several very hard classes such as bird, cat, and dog. Bottom: we illustrate segmentation examples on the MSRC-21 test set. The original image and its ground truth are shown on the left, and we plot the semantic segmentation result on the right . ***This figure is encouraged to be view in electronic version.***

and MTL-RF [9]. The data is the same as experiment I. For each test image, our approach retrieves 4 pairs of compatible-competitive references. The performance comparisons are given at the top of Fig.4, from which we clearly see that our algorithm works better than the state-of-the-arts. With the help of competitive information, we are able to distinguish the highly confusing classes (*e.g.*, cat, dog and sheep). With an unoptimized Matlab implementation, the joint segmentation task takes 6 seconds (computing features: $1s$, image retrieval: $1s$, jointly segmenting 9 images: $4s$) on a 64-bit system with Core-2 3.6 GHz CPU, 4 GB Memory. Several segmentations are illustrated at the bottom of Fig.4.

III. Size of the database. We conduct an experiment to test our approach with different sizes of training sets. The test set is the same as experiment I and II using the MSRC-21 test data. The training images come from three sources: the training data of MSRC-21, images of six classes of “building”, “grass”, “tree”, “sky”, “water”, and “road” from the Stanford Background Dataset (SBD) [4], and images of the other 15 classes from the segmentation data of PASCAL VOC 2011 [3]. The training sets of different sizes includes all the training data of MSRC-21, and their remaining images are randomly selected from SBD and PASCAL VOC 2011. For each test image, we retrieve 4 pairs of compatible-competitive references. The results are shown in Fig.3 and show that our method gets better results when the size of the database is increased.

4. CONCLUSION

In this paper a new framework is proposed to jointly segment an untagged target image with a few pairs of compatible-competitive references searched from a tagged image database. We construct cc-MIG to infer the class labels of all the superpixels of these images. The cc-MIG problem can be solved efficiently by a linear programming relaxation. Three experiments are conducted and our approach outperforms the state-of-the-arts.

5. REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? *CVPR*, 2010.

[2] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. *CVPR*, 2011.

[3] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. *http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html*, 2011.

[4] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. *ICCV*, 2009.

[5] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 2011.

[6] X. Liu, J. Feng, S. Yan, L. Lin, and H. Jin. Segment an image by looking into an image corpus. *CVPR*, 2011.

[7] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textinboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV*, 2006.

[8] J. Verbeek and B. Triggs. Region classification with markov field aspect models. *CVPR*, 2007.

[9] A. Vezhnevets and J. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. *CVPR*, 2010.

[10] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised semantic segmentation with a multi-image model. *ICCV*, 2011.