Knowledge Leverage from Contours to Bounding Boxes: A Concise Approach to Annotation

Jie-Zhi Cheng, Feng-Ju Chang, Kuang-Jui Hsu, and Yen-Yu Lin

Research Center for Information Technology Innovation, Academia Sinica, Taiwan
jzcheng@ntu.edu.tw, {fengju, kjhsu, yylin}@citi.sinica.edu.tw

Abstract. In the class based image segmentation problem, one of the major concerns is to provide large training data for learning complex graphical models. To alleviate the labeling effort, a concise annotation approach working on bounding boxes is introduced. The main idea is to leverage the knowledge learned from a few object contours for the inference of unknown contours in bounding boxes. To this end, we incorporate the bounding box prior into the concept of multiple image segmentations to generate a set of distinctive tight segments, with the condition that at least one tight segment approaching to the true object contour. A good tight segment is then selected via semi-supervised regression, which bears the augmented knowledge transferred from object contours to bounding boxes. The experimental results on the challenging Pascal VOC dataset corroborate that our new annotation method can potentially replace the manual annotations.

1 Introduction

Class based image segmentation [1–6] is the task of labeling pixels with several predefined object classes or background in an image. Distinct from the image driven segmentation task e.g., [7–11], class based image segmentation aims to not only identify the object classes of interest, but also determine the shapes or boundaries of these objects. Namely, it de facto involves in resolving two of the most fundamental problems in vision research: recognition and segmentation. Accordingly, it plays an essential role in many high-level computer vision applications, such as image and scene understanding.

Recently, significant progress for addressing the class based image segmentation has been made with the advances in many aspects, such as designing powerful visual features [1, 12], fusing information from various ways of image quantization [2, 4, 5], or exploring contextual relations between object classes [3, 6]. These approaches are implemented upon graphical models, especially *conditional random fields (CRFs)* [13], for the expressive power of modeling diverse cues and enforcing spatial consistency. However, learning graphical models in these approaches typically relies on a sufficient number of training data in the form of object *contours*. In general, the object contours are manually drawn or delineated by tools with intensive user interaction. Since learning complex graphical models typically requires large training data, the labeling cost of training data deems to be one of the major concerns for class based segmentation.

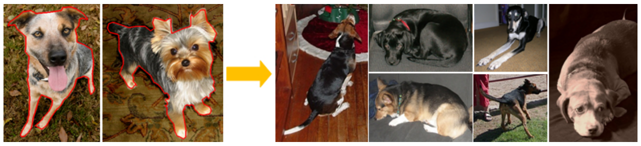


Fig. 1. Knowledge transfer from contours (*left*) to infer the unknown object contours enclosed by bounding boxes (*right*).

In this work, we introduce a concise annotation method to collect training data for class based image segmentation. Specifically, the annotation can be done with the drawing of bounding boxes. The bounding box annotation is pretty simple since we only have to click the four outer most boundary points of the object. Fig. 1 illustrates the problem setting of this study. Given a few contours as well as a set of bounding boxes of an object class, we would like to transfer the knowledge carried by the few contours to the bounding boxes. With the transferred knowledge, the object contour enclosed with the bounding box will be inferred as a training instance for the task of class based image segmentation. This work distinguishes itself with the following three main contributions.

First, we integrate the *bounding box prior* [14] into the concept of *multiple image segmentations* [2, 15, 16] as a new algorithm that automatically generates a set of *tight segments* [14] for each bounding box, and at least one of these tight segments would be close to the ground truth. An example of bounding box and its tight segments yielded by our approach are shown in Fig. 2. In this way, the task of figure-ground segmentation within this bounding box can be achieved by picking the best tight segment from the generated ones.

Second, we cast the tight segment selection for bounding boxes of an object class as a *semi-supervised regression* problem. Suppose that we are given a set of bounding boxes, and a few of them come with the object contours. In the regression problem, tight segments yielded from bounding boxes with ground truth serve as labeled training instances, while their target values for regression are set to reflect how well these segments approach the ground truth. As for tight segments without ground truth, we derive a difference upper bound of the target values of each segment pair in a bounding box. These bounds are formulated as additional constraints to regularize the learning process of the regressor. It alleviates the high risk of overfitting caused by the lack of labeled training instances. Once the regressor is obtained, the tight segment with the highest target value is selected for the bounding box.

The third contribution consists in the experiments conducted to demonstrate that our approach provides an effective alternate for manually labeled contours. We separately use the object contours obtained by manual drawings and the tight segments of bounding boxes picked by our approach as the training data for class based segmentation. Two state-of-the-art segmentation algorithms, i.e., [4, 5], are involved to compare the performances obtained by the two different sets of training data. The experimental results by each algorithm show that similar accuracy rates are achieved with either manual drawings or bounding box annotations on *PASCAL VOC* 2007 [17] segmentation task. It implies that the introduced annotation method can replace the tedious manual drawings.

2 Related Work

The literature of image segmentation is quite extensive, so our survey focuses on the key concepts relevant to the establishment of the proposed framework.

Class Based Image Segmentation. Approaches of this category, e.g., [1–6], aim to perform multi-class object recognition and segmentation simultaneously. Most of these segmentation approaches are established upon CRFs, since CRFs provide desirable abilities to concisely express the dependencies among random variables and observations, and enforce the consistency of labeling. For instance, Shotton et al. [1] propose a rich set of features to capture the texture, layout and contextual information of object classes in pixel level, and combine these features via solving an energy minimization problem over CRFs. Kohli et al. [3] and Gonfaus et al. [6] model the interaction between object classes by incorporating higher order potential functions into CRFs. Ladický et al. [4] integrate features extracted from different levels of image quantization by developing a hierarchical generalization of CRFs. Despite the effectiveness of these work, the annotation bottleneck for compiling sufficient training data remains unsolved.

Figure-Ground Segmentation. Some notable methods of this category, such as *graph-cut* [8], *GrabCut* [11], *constrained parametric min-cuts* [16], cast this task as an energy minimization problem over graph structures. A latter improvement of GrabCut is made by Lempitsky et al. [14] with the so-called *bounding box prior*. They show that the resulting foreground regions are sufficiently tight with respect to the given bounding boxes. Instead of working on individual images, the authors of [18–20] extend figure-ground segmentation for a set of images of an object class. This way, additional class-specific cues can be included to benefit figure-ground segmentation. Due to the inherent difficulty of unsupervised segmentation, the steps of segmenting objects and learning class models in [18–20] are carried out either alternately or sequentially. However, segmentation methods being aware of object classes may suffer from the problems caused by large intra-class variations or partial occlusions.

Multiple Image Segmentations. Classic image based segmentation methods, such as *normalized cuts* [7] or *mean-shift* [9], are developed with theoretic support. Nevertheless, the general conclusion [21] is still that the resulting segmentations typically are not good enough for discovering object contours. Since there is barely universal single-shot solution or parameter setting to segment out various objects with satisfactory results, the strategy of multiple image segmentations, e.g., [2, 15, 16, 22, 23], arises, in which many segmentations are computed with different segmentation algorithms, parameter settings, and/or seeds. In [15, 16], the authors assume that each object can be discovered by at least one segment. In [2], Pantofaru et al. propose to seek the most probable objects based on the intersections of multiple segments. Distinct from these approaches, we are motivated by the fact that the bounding box of an object can be acquired with low labeling cost (four clicks) but contains rich information for object inference. We couple the concepts of the bounding box prior and multiple image segmentations into a framework to estimate the object segments enclosed in the bounding boxes.

3 Inferring Multiple Tight Segments in a Bounding Box

In this section, we present an algorithm that automatically generates a set of *tight segments* for the bounding box of an object, and at least one of these tight segments would approach the object segment. Our goal in this step is to account for the information asymmetry between an object segment and its bounding box, since the latter can be determined once the former is given, but not vice versa. Specifically, we model the ambiguity in inferring the object segment from a bounding box by generating multiple segment hypotheses. If at least one of them is close to the object segment, the underlying task of inferring the object segment from a bounding box is reduced to a segment selection problem.

In the following, the approach by Lempitsky et al. [14] that yields one tight segment for a given bounding box is first reviewed. We then specify how to generalize their approach to obtain a few tight segments and make sure that at least one of them approaches the object contour.

3.1 Tight Segment via Bounding Box Prior

Let us consider a bounding box $\mathcal{I}$ of an object segment. We start by partitioning $\mathcal{I}$ into *superpixels* by mean-shift [9], which attains a fast and stable over-segmentation. In practice, the bandwidth parameters in mean-shift algorithm are adjusted by binary search, so that about 50 superpixels are obtained. Let $\mathcal{B}$ denote the set of the superpixels. A figure-ground segmentation or a segment can then be represented by a labeling vector $\boldsymbol{\ell} = [l_p] \in \{0,1\}^{|\mathcal{B}|}$, where l_p takes the value 1 if superpixel p belongs to foreground, otherwise 0.

We are particularly interested in *tight segments* within bounding box $\mathcal{I}$. Here a segment is tight with respect to $\mathcal{I}$ if the smallest rectangle covering this segment is $\mathcal{I}$ itself. It is obvious that any non-tight segments won't be the object segment. In [14], Lempitsky et al. introduce the *crossing paths* of a bounding box, and prove that a segment is tight if and only if it intersects all the crossing paths. It turns out that a tight segment $\boldsymbol{\ell}$ can be obtained by solving

$$\min_{\boldsymbol{\ell}} \sum_{p \in \mathcal{B}} U_p \cdot l_p + \lambda \sum_{(p,q) \in \mathcal{E}} V_{p,q} \cdot |l_p - l_q| \tag{1}$$

$$\text{subject to } \forall p \quad l_p \in \{0,1\}, \tag{2}$$

$$\forall C \in \mathit{\Gamma} \quad \sum_{p \in C} l_p \geq 1, \tag{3}$$

where $\mathcal{E}$ is the set of pairs of adjacent superpixels. The *unary potential* U_p specifies the preference of assigning superpixel p to either foreground or background. The *pairwise potential* $V_{p,q}$ ensures the smoothness between superpixel p and q. The nonnegative coefficient λ controls the importance tradeoff between the unary and pairwise terms. $\mathit{\Gamma}$ is the set of all the crossing paths of $\mathcal{I}$.

Note that the constraints (3) cause that the energy minimization problem (1) can no longer be solved by an efficient algorithm, like graph-cut [8]. Thus Lempitsky et al. instead solve a series of its linear relaxation, in which active constraints in (3) are added incrementally.

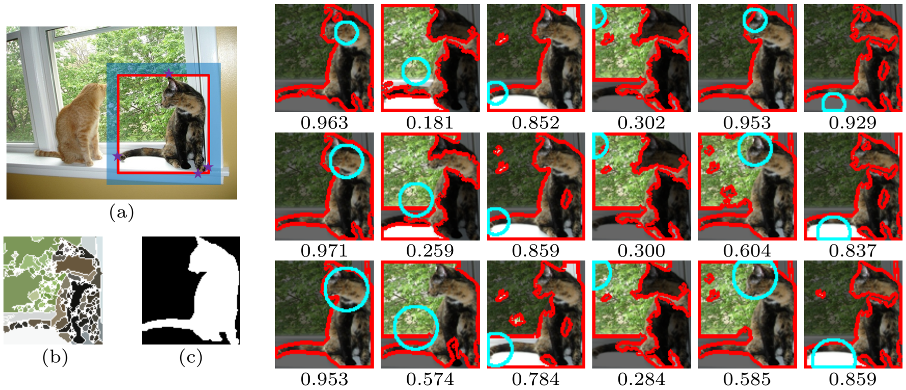


Fig. 2. (a) A bounding box defined by the four clicks (*purple stars*) for a kitty. Background seeds are placed in the blue highlighted region. (b) The superpixels of the bounding box. (c) The object segment (*ground truth*). (*Rest*) A few tight segments, marked by red contours, together with their accuracy by our approach. Each of them is generated with its respective seed region for foreground (cyan circle). These regions are sampled with different locations (*columns*) and radii (*rows*).

3.2 Multiple Tight Segments

The resulting segment by solving Eq.(1) is tightly enclosed by the given bounding box, and hence the aspect ratio of the object is maintained. Due to the unsupervised nature, a satisfactory figure-ground segmentation is not always guaranteed in our empirical test. When addressing bounding boxes of objects with multimodal color distributions and/or with clutter background, this shortcoming becomes even more evident. Alas, it is usually the case in nowadays benchmark databases of object segmentation, like MSRC-21 [1] or Pascal VOC [17].

We resolve this difficulty by implementing multi-segmentation relaxation. Namely, we generate a few tight segments with different *seeds* [16, 22], and relax the requirement to that at least one of them closely approaches the unknown object segment. It can be observed that apart from the property of tightness, the bounding box of an object also gives two additional hints for discovering the object segment: (1) Its outside borders provide strong cues for identifying the background in the bounding box; (2) It exists a few ROIs that are fully filled by the foreground. If we can retrieve one of them, it helps much in revealing the object segment. Specifically, we maintain the aspect ratio and expand the bounding box by 10%. The *background seeds* are the pixels outside the bounding box and inside the expanded one, i.e., those in the blue highlighted region in Fig. 2(a). We sample multiple sets of *foreground seeds* to account for the uncertainty on the locations and scales of those ROIs fully filled by the object. One circular seed region for foreground is constructed for the centroid of each superpixel and with each of predefined radii. The cyan circles in Fig. 2 show some of the seed regions for foreground.

We leverage the flexibility in developing potential functions $\{U_p\}$ and $\{V_{p,q}\}$ in Eq.(1), and derive one tight segment for each set of foreground seeds. A Gaussian mixture model GMM_f with five components is learned with the foreground seeds in RGB color space. Similarly GMM_b is acquired with the background seeds. For each superpixel p, the unary potential U_p is defined as

$$U_p = \sum_{u \in p} \log P(c_u|GMM_b) - \log P(c_u|GMM_f), \tag{4}$$

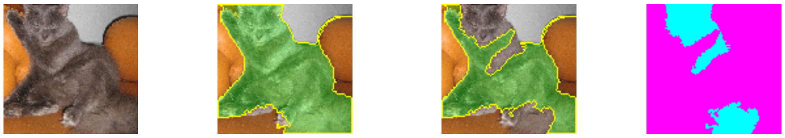
where u is an image pixel and c_u is its RGB color vector. On the other hand, the pairwise potential $V_{p,q}$ between superpixels p and q is given by

$$V_{p,q} = \sum_{u \in p, v \in q, (u,v) \in \mathcal{N}} \frac{1}{dist(u,v)} \cdot \exp{(-\beta||c_u - c_v||^2)}, \tag{5}$$

where $\mathcal{N}$ is set of neighboring pixels. We use 8-connected neighbors, and $dist(u,v)$ is the Euclidean distance between pixels u and v. β is a positive constant. One tight segment is inferred by optimizing Eq.(1) with these redefined potentials in Eq.(4) and Eq.(5). The procedure is repeated for each combination of foreground seed regions and parameter settings (λ in Eq.(1) and β in Eq.(5)). Multiple tight segments of the bounding box are then produced.

An example is shown in Fig. 2. The left three figures give the bounding box, its representation in superpixels, and the ground truth (GT) respectively. The others are 18 of the yielded tight segments for the bounding box. We evaluate the goodness of a segment, say $\boldsymbol{\ell}$, by $1 - \frac{XOR(R(\boldsymbol{\ell}),GT)}{\#pixel}$, where XOR is the function of *exclusive or*, and $R(\boldsymbol{\ell})$ is a binary vector that indicates each pixel in $\boldsymbol{\ell}$ assigned to either foreground or background. Hereafter we will use the pixelwise XOR function to measure the goodness of a segment w.r.t. the ground truth, or the overlapping between two segments. From Fig. 2, it can be observed that seed regions for foreground located within the object and with proper radii often lead to satisfactory tight segments. Since the object must appears in some location of the bounding box with one particular scale, the seeding strategy with high chance will discover at least one tight segment close to the ground truth.

Redundance Removal. Each generated tight segment is parameterized by the location and scale of the seed region for foreground, and the values of λ and β. In our implementation, the number of the tight segments generated for a bounding box is in the order of 10^3. Since many of them are redundant, we develop a $(1-\epsilon)$-approximation procedure to compile the tight segmentations into a smaller set of representative ones. In initialization, all the tight segments are sorted in a queue according to their scores measured by *ratio cut* [24]. We *pop* the first tight segment, add it into the representative set, and remove all the tight segments of more than $1-\epsilon$ overlapping with it from the queue. The process is done repeatedly until the queue is empty. It is obvious that the best tight segment remained in the representative set shares at least $1-\epsilon$ overlapping with the original best one. We empirically set ϵ as 0.05 in all the experiments.



(a) bounding box (b) tight segment A (c) tight segment B (d) XOR(A,B).

Fig. 3. For each pair of tight segments in a bounding box, an upper bound on the difference of their target values can be determined. See text for the details

4 Semi-supervised Regression for Segment Selection

Given a few contours as well as a set of bounding boxes of an object class, we illustrate how to infer the object segments of these bounding boxes by solving a semi-supervised regression problem in this section.

4.1 Our Formulation

Consider a bounding box set D, which is collected from object segments of a class. A few bounding boxes in D come with the object contours (ground truth), i.e., $D = L \cup U$, where $L = \{B_i, GT_i\}_{i=1}^{\ell}$, $U = \{B_i\}_{i=\ell+1}^{\ell+u}$, and $\ell << u$. We generate multiple tight segments for each bounding box B_i by the procedure described above. That is, $B_i = \{\mathbf{x}_{ij}\}_{j=1}^{N_i}$, where N_i is number of the yielded tight segments, and $\mathbf{x}_{ij}$ is the feature vector of the jth tight segment. Our goal is to infer the object segments of these bounding boxes. Since at least one tight segment with high probability is close to the object contour, this goal can be accomplished by picking the tight segment as close to object contour possible. We cast this task as a semi-supervised regression problem.

We start by creating the *labeled* training instances for the regression problem. Inspired by work [16, 22, 23] of ranking multiple segmentations or proposals, we treat each tight segment in a bounding box with the ground truth as one labeled instance, whose *target value* is set via computing the pixelwise XOR function w.r.t. the ground truth as mentioned before. A set of labeled training instances is then produced, i.e., $\{(\mathbf{x}_{ij}, y_{ij})\}_{(i,j)\in S_L}$, where $S_L = \{(i,j)|1 \le i \le \ell, 1 \le j \le N_i\}$. Unlike [16, 22, 23] where sufficient training instances are available from other sources, we have too few labeled bounding boxes to stably derive the regressor. The unfavorable effect of *overfitting* hence may occur.

We resolve this problem by introducing the *unlabeled* tight segments. For each pair of tight segments in a bounding box, an upper bound on the difference of their target values can be derived without the ground truth. Let's illustrate it via Fig. 3. Given a pair of tight segments A and B, the bounding box can be divided into two regions according to their labeling consistence, i.e., the magenta (consistent) and cyan (inconsistent) regions in Fig. 3(d). Suppose that the inconsistent part takes $\theta \times 100\%$ area of the bounding box. It can be verified that the difference between the target values of A and B is at most θ, since the target value is defined as the percentage of *correct* pixels, and only the inconsistent region contributes to the difference of their target values. Thus a set of these bounds is yielded, i.e., $\{(\mathbf{x}_{ij}, \mathbf{x}_{ij'}, \theta_{ijj'})\}_{(i,j,j')\in S_U}$, where $S_U = \{(i,j,j')|\ell+1 \le i \le \ell+u, 1 \le j < j' \le N_i\}$, and $\theta_{ijj'}$ is the bound between segments $\mathbf{x}_{ij}$ and $\mathbf{x}_{ij'}$ of bounding box B_i.

Integrating the labeled and unlabeled tight segments, the semi-supervised regression problem is formulated as the following constrained optimization problem

$$\min_{\mathbf{w},b,\{\xi_{ij}\},\{\rho_{ijj'}\}} \quad \frac{1}{2}||\mathbf{w}||^2 + C_\ell \sum_{(i,j)\in S_L} \xi_{ij} + C_u \sum_{(i,j,j')\in S_U} \rho_{ijj'} \tag{6}$$

$$\begin{aligned} \text{subject to} \quad & \mathbf{w}^\top \mathbf{x}_{ij} + b - y_{ij} \leq \varepsilon + \xi_{ij}, \text{ for } (i,j) \in S_L, & (7)\\ & y_{ij} - \mathbf{w}^\top \mathbf{x}_{ij} - b \leq \varepsilon + \xi_{ij}, \text{ for } (i,j) \in S_L, & (8)\\ & \mathbf{w}^\top \mathbf{x}_{ij} - \mathbf{w}^\top \mathbf{x}_{ij'} \leq \theta_{ijj'} + \rho_{ijj'}, \text{ for } (i,j,j') \in S_U, & (9)\\ & \mathbf{w}^\top \mathbf{x}_{ij'} - \mathbf{w}^\top \mathbf{x}_{ij} \leq \theta_{ijj'} + \rho_{ijj'}, \text{ for } (i,j,j') \in S_U, & (10) \end{aligned}$$

where $\mathbf{w}$ and b are parameters of the learned regressor, $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$. $\{\xi_{ij}\}$ and $\{\rho_{ijj'}\}$ are two sets of slack variables that are nonnegative, and are used to measure the degrees of violation in the corresponding constraints. C_ℓ, C_u, and ε are nonnegative constants whose values are determined via cross validation.

We now justify for the above optimization problem. The first two terms in Eq.(6) together with constraints in Eq.(7) and Eq.(8) jointly lead to the formulation of *support vector regression* [25]. The constraints in Eq.(9) and Eq.(10) result from pairs of unlabeled tight segments. In other words, the regressor is derived by not only fitting the labeled segments but also preserving the implicit structure of the unlabeled segments. Despite the complexity of Eq.(6), it is a *quadratic programming* (QP) problem, and there exist efficient solvers, e.g., *MOSEK* [26], for optimization.

The optimization problem still cannot be handled by a QP solver due to the large number of constraints in Eq.(9) and Eq.(10). However, this is not as hard a problem as it may seem at the first glance, since the *active* constraints in Eq.(9) and Eq.(10) are quite sparse. It implies that we can tackle this issue via the *cutting-plane method* [27] where a *working constraint set* is maintained by adding the most violated constraints incrementally. In our case, we start with a empty working set. At each iteration, we add the most violated constraint to the working set for each bounding box, if any. Iterations are repeated until no constraints can be added or a maximum number of iterations is reached.

Once the regressor, $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$, is obtained, we infer the object segment for each bounding box B_i as the j^*th tight segment with $j^* = \arg\max_j f(\mathbf{x}_{ij})$.

4.2 The Adopted Features for Segment Description

We implement a set of *mid-level* features, suggested in [16, 23], for characterizing each tight segment, including 1) *Percentage of boundary pixels*: The ratio of the number of boundary pixels to the number of foreground pixels; 2) *Boundary edge strength*: The edge strengths along the object contour; 3) *Centroid*: The normalized coordinates of the mass center of the segment; 4) *Major and minor axis length*: The lengths of the major and minor axes of the ellipse that approximates the segment; 5) *Convexity and area*: The ratios of the number of foreground pixels to the area of the convex hull and to the whole bounding box; 6) *Foreground and background dissimilarity*: The dissimilarity is respectively measured by three visual features, i.e., color, *SIFT* [28], and *Texton* [1]. A pair of histograms, one

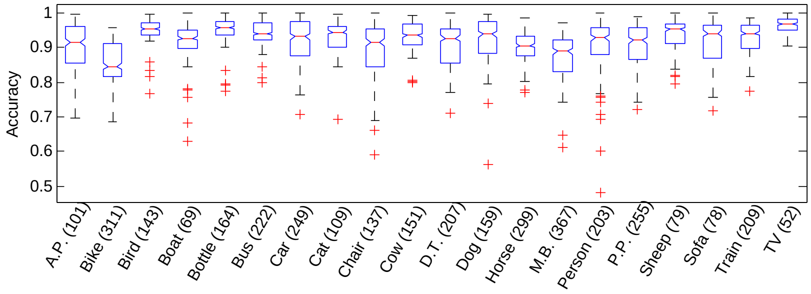


Fig. 4. Boxplot of the best accuracy rates of the best tight segments w.r.t. each of the 20 object classes. The edges of each box are the 25th and 75th percentiles. Outliers are marked as red-cross signs. The numbers of the generated tight segments are given.

for foreground and one for background, over the quantized clusters is yielded for each feature. The χ^2 distance is employed for dissimilarity measure.

5 Experimental Results

To evaluate the performances of the proposed approach, three experiments are carried out on Pascal VOC 2007, a benchmark dataset for object segmentation. We investigate the validness of the assumption that there exists at least a good one in the pool of multiple tight segments in the first experiment. The efficacy of the semi-supervised regression model for segment selection is assessed in the second experiment. We demonstrate the effectiveness of using the bounding box annotations as training data for the class based image segmentation algorithms in the last experiment.

5.1 Dataset: Pascal VOC 2007

The Pascal VOC 2007 Segmentation Challenge contains 21 categories, including 20 object classes with the plus of background. Each object category contains about 30 to 100 annotated objects, except the class of `person`, which has more than 300 ones. Due to the large intra-class variations in this dataset, the annotation cost of segmentation is conceivably substantial. It hence serves as good test beds for corroborating our purpose of concise annotations, and for justifying the effectiveness of the proposed approach. In our work, the training and validation data in this dataset are used in the experiments I and II, and the selected tight segments by the semi-supervised regression models are treated as training data in the experiment III.

5.2 Experiment I: Multiple Tight Segments

The effectiveness of multiple segmentation strategy lives with the underlying assumption that there is at least one tight segment close to the object segment. To inspect if this assumption is held in our method, we crop the bounding box of each annotated object segment in the training and validation sets, and generate a set of tight segments for the bounding box. The resulting tight segments are

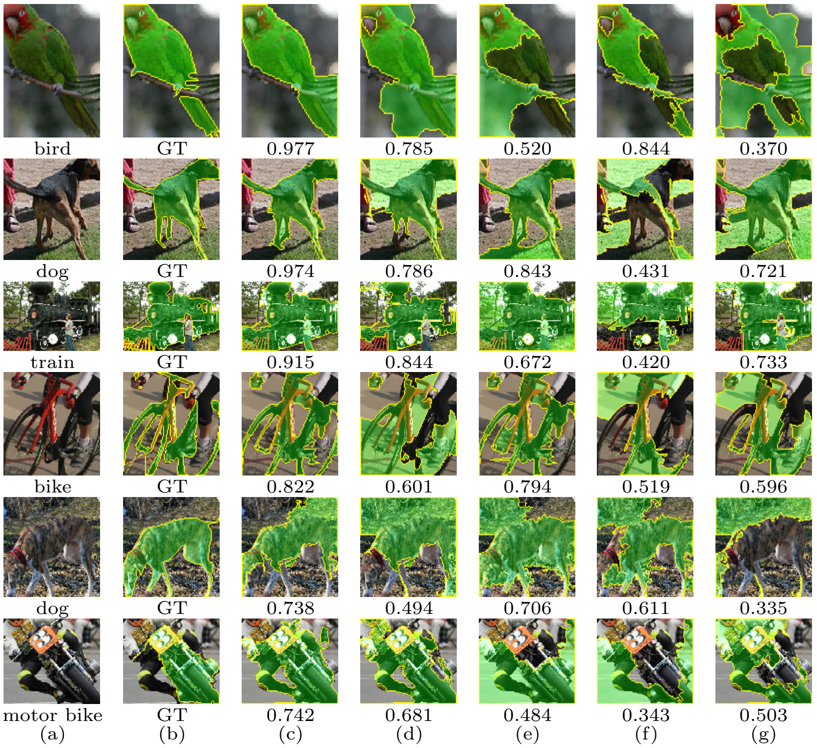


Fig. 5. Examples of the yielded multiple tight segments. (a) Bounding box. (b) Ground truth. (c) The best tight segment and its accuracy. (d) ∼ (g) Other tight segments.

compared to the annotated ground truths. Specifically, the input tight segments for the semi-supervised regression models of each class are first evaluated with the accuracy metrics. The tight segment with the highest accuracy of a bounding box is then regarded as the performance upper bound of the semi-supervised regression model. Figure. 4 depicts the `MATLAB' boxplot` of the best accuracy distributions of bounding boxes w.r.t. each of the 20 object classes. Meanwhile, the average numbers of the yielded tight segments are also reported in Fig. 4, with the variation from 52 to 367. The number of the yielded tight segments for each bounding box typically depends on the complexity of object appearance and the foreground/background discernibility. In general, it can be found that most bounding boxes hold *good* tight segments with accuracy rates higher than 0.9, except for those of class `bike`. However, the lowest Q1, i.e., the 25th percentile, of class `bike` is still higher than 0.8. This may set up a good foundation for the semi-supervised regression models to pick satisfying tight segment for the objects within the bounding boxes.

Figure. 5 lists several cases of the generated tight segments for visual assessment. The first three rows show the examples where the accuracy rates of the best tight segments are higher than 0.9. The last three rows give the cases where the proposed approach doesn't perform well, including objects constituted with fine details in the example of `bike`, objects sharing similarly color components

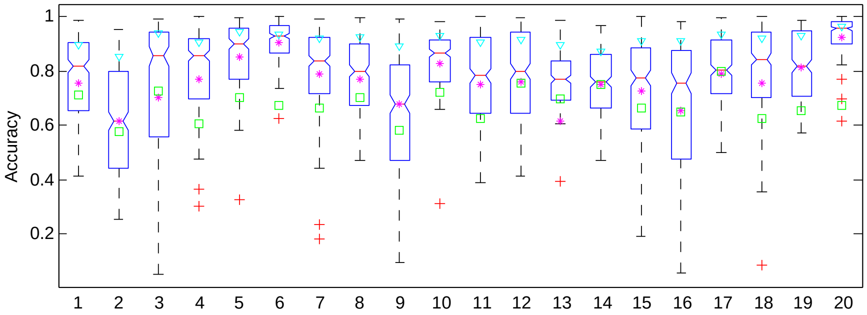


Fig. 6. The accuracy rates of the semi-supervised regressor (*plotted with* `boxplot`) are compared with those of the best tight segments (*cyan triangles*), of GrabCut (*green squares*), and of the supervised regressor (*magenta* $*$ *signs*).

with background in the latter example of `dog`, and objects co-presented with other objects in the example of `motor bike`.

5.3 Experiment II: Segment Selection for Object Contour Estimation

The experiment II is designed to assess the quality of the selected tight segments by the regressors derived via the proposed supervised (without constraints (9) and (10)) and semi-supervised regression models. Similar to the experiment I, the quality of the selected tight segment is evaluated with the accuracy metrics. To give comparative study, alternative segmentations from *GrabCut* [11] algorithm are involved in this experiment. The foreground model in GrabCut is initialized with the center 50% area of the bounding box, while the background model is fitted from the same background sample region of our method.

Both the supervised and semi-supervised models learn a regressor for each of the 20 object categories in Pascal VOC 2007. For each category, randomly selected 10% of bounding boxes come with the ground truths (object contours), while the rest are treated as unlabeled. In learning the regressors, all parameters (C_ℓ and C_u in(6), ε in (7)) are automatically determined via cross validation. The accuracy distributions of the tight segments selected by the semi-supervised regressor for each class are depicted with function `boxplot` in Fig. 6. Meanwhile, the average accuracy rates of each class w.r.t. the best tight segments in experiment I, the segments picked by supervised regressor, and the comparative segmentations from GrabCut are also plotted as the cyan triangles, magenta $*$ signs, and green squares, respectively.

It can be found in Fig. 6 that in most classes the tight segments selected by the semi-supervised regressor hold higher accuracy rates than the segments yielded by GrabCut or picked by the supervised regressor. It reveals that the additional constraints induced by unlabeled tight segments facilitate the training process, and lead to a regressor with low generalization error. The worst performance of our approach arrives at the 2nd class, i.e., `bike`. In this class, regressors

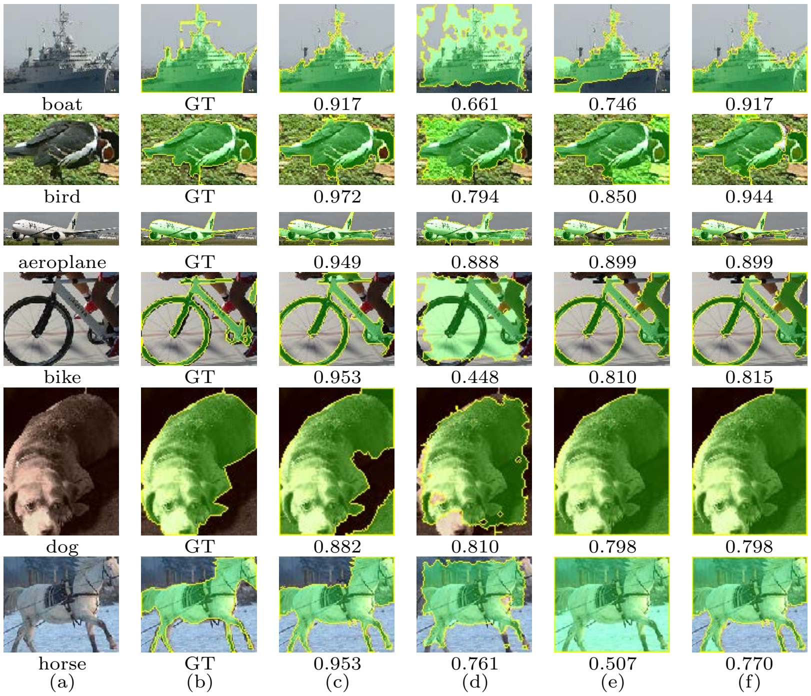


Fig. 7. Inferred object segments by various approaches. (a) Bounding box. (b) Ground truth. (c) The best tight segment. (d) Segment yielded by GrabCut. (e) Segment selected via supervised regression. (f) Segment selected via semi-supervised regression.

learned by either supervised or semi-supervised models obtain similar accuracy rates around 0.6. The determined value of C_u approaches zero in this class. It implies that satisfying the additional constraints is not helpful for reducing the validation error. We infer that it may result from the large intra-class variations.

To give visual assessment, Fig. 7 lists the tight segments selected from the supervised and semi-supervised regressors with the comparison to the ground truths, best tight segments, and the segmentation results by GrabCut.

5.4 Experiment III: Class Based Image Segmentation

The experiment III aims to verify the effectiveness of the concise annotations for the class based image segmentation methods. To this end, the tight segments selected by the supervised and semi-supervised regression models, GrabCut segments, and ground truths in the experiment II are treated as training annotations for two state-of-the-art class based segmentation methods [4, 5]. The yielded segments from the two variational regression models and GrabCut are pasted back to the original training and validation images for reasonable comparison to the ground truths. The two segmentation methods [4, 5] learnt from the distinctive four sets of annotations are further evaluated with the testing data of the PASCAL VOC 2007 Segmentation dataset.

Table 1. Quantitative results of methods [4, 5] on Pascal VOC Segmentation task w.r.t. annotations of the ground truth, GrabCut, supervised and semi-supervised regression models.

	Ground Truth	GrabCut	Supervised Reg.	Semi-Sup. Reg.
Hierarchical CRF [4]	**11.23**	9.96	11.06	10.64
CRF+$N = 0$ [5]	**14.10**	12.47	13.29	13.33
CRF+$N = 2$ [5]	25.26	24.56	**26.51**	**26.51**
CRF+$N = 4$ [5]	23.92	21.31	24.81	**24.85**

Table 1 summarizes the averaged performances of the methods [4] and [5] w.r.t. the four sets of training data. In Table 1, the performances of the method [4] are first given, while the performances of the method [5] under the three settings, i.e., "CRF+N=0", "CRF+N=2", and "CRF+N=4", are then reported respectively. Variable N here indicates the neighborhood size in [5]. Referring to Table 1, the class based segmentations learnt from the two variations of our annotations achieve similar performances to the results of ground truth annotations. This promises our goal of using bounding boxes as concise annotations. The reason why the method [5] outperform to the method [4] may lie in that the hierarchical conditional random field model may highly rely on large training data to learn a powerful segmentation model.

Furthermore, it can be observed in the third and fourth rows of Table 1 that the accuracy rates with our annotations are slightly higher than the rates with ground truths. It may be because that the method [5] tends to overfit the difficult/noisy training data like the second dog case in Fig. 2, provided with precise annotations of manual drawings. Vague annotations resulted from our method may instead lower down the importance of this kinds of difficult training data, and lead to a surprisingly better performance.

6 Conclusions

A new concise annotation method for the task of class based image segmentation is introduced in this study. Guided by the bounding box prior, the proposed method first renders distinctive tight segments, and a good tight segment is further inferred by the supervised and semi-supervised regression models. The inferred tight segment from the given bounding box serves as a hidden object contour to train a complex graphical model for the class based image segmentation task. The results of three extensive experiments support the efficacy of our method. In the future, we will extend our inference model to account for the interaction between different classes, as the co-presentation of objects with different classes is the major limitation of our method. Moreover, performance evaluation on other challenging datasets is planned in our future work.

Acknowledgement. This study is conducted under the "III Innovative and Prospective Technologies Project" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China. It is also supported in part by grant NSC 101-2221-E-001-018.

References

1. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. In: IJCV. (2009)
2. Pantofaru, C., Schmid, C., Hebert, M.: Object recognition by integrating multiple image segmentations. In: ECCV. (2008)
3. Kohli, P., Ladický, L., Torr, P.: Robust higher order potentials for enforcing label consistency. IJCV (2009)
4. Ladický, L., Russell, C., Kohli, P., Torr, P.: Associative hierarchical CRFs for object class image segmentation. In: ICCV. (2009)
5. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: ICCV. (2009)
6. Gonfaus, J., Boix, X., van de Weijer, J., Bagdanov, A., Serrat, J., González, J.: Harmony potentials for joint classification and segmentation. In: CVPR. (2010)
7. Shi, J., Malik, J.: Normalized cuts and image segmentation. TPAMI (2000)
8. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. TPAMI (2001)
9. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. TPAMI (2002)
10. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV (2004)
11. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": Interactive foreground extraction using iterated graph cuts. In: SIGGRAPH. (2004)
12. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: CVPR. (2008)
13. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. (2001)
14. Lempitsky, V., Kohli, P., Rother, C., Sharp, T.: Image segmentation with a bounding box prior. In: ICCV. (2009)
15. Galleguillos, C., Babenko, B., Rabinovich, A., Belongie, S.: Weakly supervised object localization with stable segmentations. In: ECCV. (2008)
16. Carreira, J., Sminchisescu, C.: Constrained parametric min-cuts for automatic object segmentation. In: CVPR. (2010)
17. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: (The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results)
18. Winn, J., Jojic, N.: LOCUS: Learning object classes with unsupervised segmentation. In: ICCV. (2005)
19. Cour, T., Shi, J.: Recognizing objects by piecing together the segmentation puzzle. In: CVPR. (2007)
20. Alexe, B., Deselaers, T., Ferrari, V.: ClassCut for unsupervised class segmentation. In: ECCV. (2010)
21. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. TPAMI (2007)
22. Endres, I., Hoiem, D.: Category independent object proposals. In: ECCV. (2010)
23. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: CVPR. (2011)
24. Wang, S., Siskind, J.: Image segmentation with ratio cut. TPAMI (2003)
25. Vapnik, V.: Statistical Learning Theory. Wiley (1998)
26. The MOSEK Optimization Software: (http://www.mosek.com/index.html)
27. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural SVMs. ML (2009)
28. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)