CSI: Composite Statistical Inference Techniques for Semantic Segmentation

Fuxin Li FLI@CC.GATECH.EDU
Georgia Institute of Technology
Joao Carreira JOAOLUIS@ISR.UC.PT
ISR - University of Coimbra
Guy Lebanon LEBANON@CC.GATECH.EDU
Georgia Institute of Technology
Cristian Sminchisescu CRISTIAN.SMINCHISESCU@MATH.LTH.SE
Lund University

Abstract

We present an inference procedure for the semantic segmentation of images , based on estimates of the overlap between each of a set of large object segmentation proposals and the objects present in the image. We extend the composite likelihood methodology to error distributions on such one-dimensional statistical estimates, and define continuous latent variables on superpixels obtained by multiple intersections of segments, then output the optimal segments from the inferred superpixel statistics. The algorithm is capable of *recombine* initial mid-level proposals, as well as *handle multiple interacting objects, even from the same class*, all in a consistent joint inference framework by maximizing the composite likelihood of the underlying statistical model using an EM algorithm. In the PASCAL VOC segmentation challenge, the proposed approach obtains high accuracy and successfully handles images with complex object interactions.

1. Introduction

Semantic segmentation aims to detect objects from different categories and identify their spatial layout simultaneously. Each pixel in the image must be classified as a foreground object of a certain category, or be assigned as background. Because it can identify object boundaries, this challenging task is of great practical importance for scene understanding and robot vision (Arbelaez et al., 2012; Ion et al., 2011; Gould et al., 2009; Kumar et al., 2011; Ladicky et al., 2010; Pantofaru et al., 2008).

International Conference on Machine Learning (ICML) workshop on Inferning: Interactions between Inference and Learning, Atlanta, Georgia, USA, 2013.

An approach that we have pursued with some success was based on 'sliding segments', starting from an unsupervised generation of many possibly conflicting mid-level figure-ground object segmentation proposals with large spatial support, obtained based on cuts in graphs defined on edge and color potentials. The segments are then passed to classifiers or regressors that determine to which category they belong. Full image interpretations are then assembled sequentially from individual segments (Fig. 1).

The existence of predictions for many mutually overlapping segments poses a new inference challenge for pixel labeling. High-order CRF models (Ladicky et al., 2009; 2010; Dann et al., 2012) usually involve different types of pairwise potentials (between pixels, between pixels and segments and between segments) and inference can be intractable. Other approaches search for configurations of non-overlapping segment hypotheses (Gould et al., 2009; Kumar et al., 2011) by using non-maxima suppression and maximum clique random field models (Ion et al., 2011). They can be tractable since the decision space is limited to the initial segments (normally < 200 in practice). However, segments often occlude and cut through each other and the initial mid-level proposals may not be entirely accurate. Non-probabilistic methods have also been developed to produce an average (Hoiem et al., 2007; Pantofaru et al., 2008) or weighted average (Carreira & Sminchisescu, 2012; Li et al., 2010; Arbelaez et al., 2012) of the predicted scores on each pixel/superpixel, then output the highest scoring labels. This strategy would typically allow for the refinement of a semantic segmentation, but in a heuristic manner, by e.g. thresholding pixel or superpixel scores.

In this paper, we propose a model that allows for the refinement and recombination of initial bottom-up proposals using a principled statistical inference method, while avoiding some of the intractability with random field structures. Instead of directly modeling the high-dimensional condi-

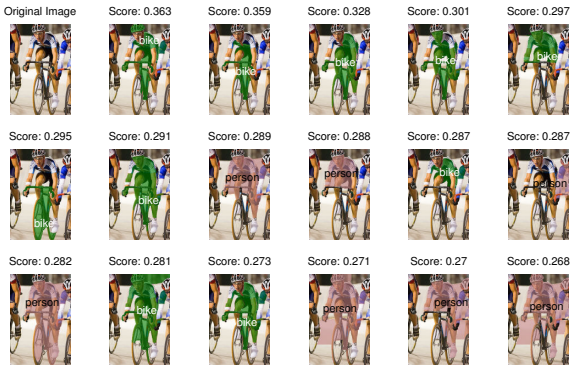

Figure 1. (Best viewed in color). The need for an efficient inference procedure given multiple object segmentation proposals. Simply performing non-maximum suppression would discard all the `person` segments, which have lower scores because they all overlap the first `bike` segment.

tional label distributions, we model the one-dimensional error distributions of many predicted *region statistics*. By combining thousands of pixels that span a large segment into one segment statistic, we avoid difficult maximum a posteriori inference in random field models with cyclic dependencies. Models of error distributions are commonly seen in the context of regression, the simplest being the Gaussian error used in least squares. In our case, the error distribution is modeled as a mixture with two components based on exploratory data analysis. The first component corresponds to false positive detections while the second one is a Gaussian truncated to the domain of the statistic.

Our main idea is to model the segments as *computable* statistics from superpixels that have no spatial overlap, which means there exists a mathematical formula that can output segment statistics given values of the superpixel statistics. From this link, we can optimize the superpixel statistics by maximizing the composite likelihood (or posterior) of the predicted segment statistics in the modeled error distribution. Intuitively, the configuration of superpixels that can explain most segment statistics will emerge as the maximum likelihood solution, as shown in fig. 2. The conceptual graphical model is presented in fig. 3.

Our methodology consists of a training phase and an inference phase. During training, regressors are estimated to predict segment statistics. This can be done by standard routines such as SVR or least-squares, and is not covered. Given a test image, the inference has three main stages:

- Use the regressors to predict segment statistics.
- Maximize the composite likelihood to estimate superpixel statistics.
- Output an optimal full-image semantic segmentation given the estimated superpixel-level statistics.

The first stage is straightforward thus we will be mainly discussing the second and the third ones.

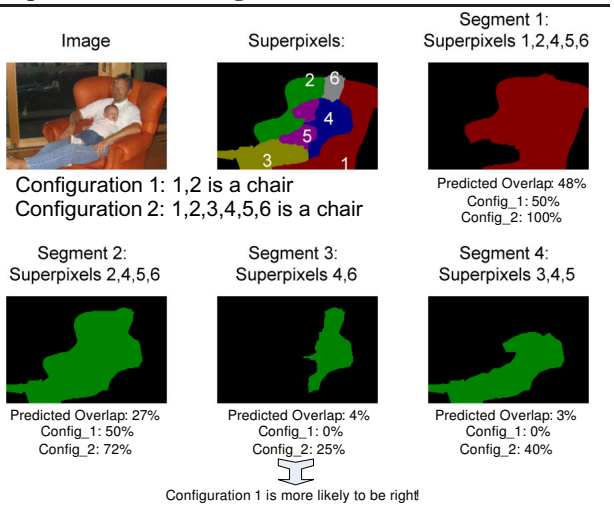

Figure 2. (Best viewed in color) The goal of our inference is to find the superpixel configuration which best explains most of the predicted segment statistics, here spatial overlap (with the chair object). This formulation allows discovering objects that are cut into disconnected components, such as the chair in this figure.

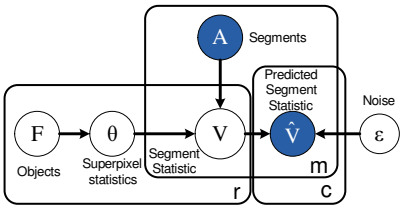

Figure 3. The conceptual graphical model. Superpixel statistics are generated from the ground truth objects. Segment statistics are generated from superpixel statistics and the segments. The observations are predicted segment statistics on each category, perturbed with noise ϵ. During inference, we first solve for the superpixel statistics θ, then output full objects given θ.

2. Composite Likelihood and Statistical Estimates

We denote $p(x)$ the probability of random variable x, $\mathbb{I}$ the indicator function. $\mathcal{N}(x; \mu, \sigma^2)$ the normal distribution with mean μ and variance σ^2, $Ber(\alpha)$ a Bernoulli distribution with parameter α, $\text{Exp}(x; \lambda)$ an exponential distribution with parameter λ, and $\delta(x)$ the Dirac function. When x is a vector, $x \geq 0$ means that all dimensions of x are larger or equal to 0. For a set A, let $|A|$ denote its cardinality. A segment is considered a set whose cardinality is the number of pixels inside it.

A maximum composite likelihood (MCL) approach (Lindsay, 1988; Dillon & Lebanon, 2010) drops the independence assumptions typical in maximum likelihood. For us, this is important, in order to be able to leverage overlapping higher-order observations (on segments) that are strongly inter-dependent.

Definition 1 *Suppose we have a dataset* D $=$

$\{X^{(1)}, \ldots, X^{(n)}\}$, where each $X^{(i)}$ is a m-dimensional vector. Consider a finite sequence of variable subset pairs (called m-pairs) $(A_1, B_1), \ldots, (A_k, B_k)$, where $A_j, B_j \subset \{1, \ldots, m\}, \forall j \in 1, \ldots, k$ with $A \neq \emptyset = A \cap B$. Given vector $\beta \geq 0$, the composite likelihood object is

$$cl(\theta) = \sum_{j=1}^{k} \beta_j \log p_\theta(X_{A_j} | X_{B_j}). \quad (1)$$

MCL solves for θ by maximizing the composite likelihood (1). It is asymptotically consistent given an identifiability assumption, but often intractable because of the need to model a high-dimensional distribution $p_\theta(X_{A_j}|X_{B_j})$. We propose to extend the MCL framework to distributions on statistical estimates, which are 1-dimensional distributions hence easier to model and estimate.

Definition 2 *With the same conditions as in Definition 1 for $D, X^{(i)}, A_j, B_j$ and β, let us further assume that $f(X^{(i)}, A_j, B_j)$ is an observed statistic from $X^{(i)}, A_j$ and B_j. We define the maximum composite f-likelihood problem as*

$$\max_\theta \sum_{i=1}^{n} \sum_{j=1}^{k} \beta_j \log p_\theta(f(X^{(i)}, A_j, B_j)). \quad (2)$$

This new MCL problem recovers the model parameters θ from the composite f-likelihood $\log p_\theta(f(X^{(i)}, A_j, B_j))$ for all the random variables on multiple different subsets. As an example, suppose a fixed-length feature vector Z_{ij} can be extracted from $X^{(i)}, A_j, B_j$ and the distribution of f_{ij} can be modeled as $p_\theta(f_{ij}) = \mathcal{N}(\theta^\top Z_{ij}, \sigma^2)$, with θ the regression weights. The MCL problem in this case becomes a weighted least squares regression of solving for θ. The asymptotic consistency proof still partially holds (see associated technical report (Li et al., 2013)), even when different f_{ij} are inter-dependent. Intuitively, as the number of observations goes to infinity, the true model parameters θ should give the best performance for each individual segment, hence converge to the optimal solution of the MCL problem (2), given a suitably chosen β vector.

3. Maximizing the Composite Likelihood for Semantic Segmentation

In this section we present the probabilistic model of the proposed CSI (Composite Statistical Inference) method that uses the modified MCL to infer superpixel statistics. The final stage output final segmentations given the superpixel statistics is discussed in sec. 4.

3.1. Semantic Segmentation from Figure-Ground

In our problem setting, I represents the image, as a lattice of pixels. An *object segmentation proposal* (or simply *segment*) $A_i \subset I$ is a subset of I. Suppose m segments $A_1, A_2, \ldots, A_m$; c object categories $C_1, C_2, \ldots, C_c$; r ground truth objects $F_1, \ldots, F_r$ are present in the image I and each one belongs in a particular category, denoted as $F_k \in C_j$. Each pixel p in the image should either belong to a single object or to the background, i.e. $\sum_{k=1}^{r} \mathbb{I}(p \in F_k) \leq 1$. For each segment A_i, its class-specific overlap with a category C_k is defined by

$$V_{ik}^0 = V(C_k, A_i) = \max_{F_j \in C_k} \frac{|F_j \cap A_i|}{|F_j \cup A_i|}. \quad (3)$$

V_{ik}^0 can be estimated by training one regressor for each category $C_k$on a separate training set (for details on possible training methods one can consult e.g. (Li et al., 2010; Carreira & Sminchisescu, 2012; Arbelaez et al., 2012)). We denote the estimates as $\hat{V}_{ik}^0$.

Given segments $A_1, A_2, \ldots, A_m$, we find *multiple intersections* by dividing the image I into superpixels $S_1, S_2, \ldots, S_n$, so that $\forall i, j, S_i \cap S_j = \emptyset, \forall k, A_k = \cup_i S_{k(i)}$ (every segment A_k is the union of some superpixels), and the number of superpixels is minimal. In practice we consider only segments that have non-negligible predicted overlap (over a loose threshold) with at least one category. Therefore, in many cases, the superpixels have finer granularity inside objects of interest (fig. 4) and coarser granularity on the background.

3.2. The Probabilistic Model

We use θ_{kj} to model the percentage of pixels within a superpixel S_k that belongs to object F_j. Then, the overlap between a segment A_i and F_j can be computed as

$$V_{ij} = \frac{|F_j \cap A_i|}{|F_j \cup A_i|} = \frac{\sum_{S_k \in A_i} \theta_{kj} |S_k|}{\sum_{S_k \in A_i} |S_k| + \sum_{S_k \notin A_i} \theta_{kj} |S_k|} \quad (4)$$

Importantly, V_{ij} is computable from θ only since each $|S_k|$ is a constant. Now, given the observed overlaps $\hat{V}_{ij}^0$, one can optimize θ by maximizing the composite likelihood of $\hat{V}_{ij}^0$, given the overlap $V_{ij}(\theta)$ computed from θ:

$$\max_\theta \sum_{i=1}^{m} \sum_{k=1}^{c} \max_{F_j \in C_k} \log p(\hat{V}_{ik}^0 | V_{ij}(\theta)) \quad (5)$$

where the inside $\max$ operation represents the fact that $\hat{V}_{ik}^0$ is an estimate of $\max_{F_j \in C_k} V_{ij}(\theta)$, instead of any $V_{ij}(\theta)$. If we know the number of objects in each category and their rough locations, this can be solved by assigning each $\hat{V}_{ik}^0$ to one of the objects in C_k, so that likelihood is maximized (see (Li et al., 2013) for details). After the assignment, suppose each $\hat{V}_{ij}$ has been properly assigned from a corresponding $\hat{V}_{ik}^0$, if $F_j \in C_k$. The MCL problem becomes:

$$\max_\theta \sum_{i=1}^{m} \sum_{j=1}^{r} \beta_{ij} p(\hat{V}_{ij} | V_{ij}(\theta)) \quad (6)$$

where θ is an $n \times r$ matrix, $\beta_{ij} = 1$ if segment A_i has been assigned to object F_j and 0 otherwise.

Fig. 5 shows histograms on $V|\hat{V}$, for the data collected from PASCAL VOC training set, which can easily be interpreted as a combination of two components: a bump at $V = 0$, which apparently corresponds to false positive detections, and a centered distribution with $V \neq 0$. As $\hat{V}$ increases, the chance of misclassification is reduced.

Motivated by these observations, we introduce an additional Bernoulli random variable z_{ij} for each predicted score $\hat{V}_{ij}$ (fig. 6). The outcome of z_{ij} informs whether the prediction $\hat{V}_{ij}$ is a false positive. We make three conditional distribution assumptions:

$$\begin{aligned} V_{ij}|\hat{V}_{ij}, z_{ij} &\sim \begin{cases} Exp(\lambda), & z_{ij} = 0 \\ \mathcal{N}(\hat{V}_{ij}, \sigma^2), & z_{ij} = 1 \end{cases} \\ z_{ij}|\hat{V}_{ij} &\sim Ber(\alpha(\hat{V}_{ij})) \\ z_{ij} = 1|\hat{V}_{ij}, V_{ij}, \theta &\sim \Pr(z_{ij} = 1|V_{ij}, \hat{V}_{ij}) f(V_{ij}, \theta_{-j}) \end{aligned} \tag{7}$$

where $\theta_{-j} = [\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_r]$ represents all the θ columns without the j-th. The assumptions are in line with our observations: if $z_{ij} = 0$, the prediction is a false positive and the overlap V_{ij} should be 0. We take an exponential distribution as an approximation, due to smoothness and tractability. If $z_{ij} = 1$, then V_{ij} should be centered around the predicted overlap[1]. Besides, the false positive probability $p(z_{ij} = 0|\hat{V}_{ij})$ controlled by $\alpha(\hat{V}_{ij})$ is smaller if $\hat{V}_{ij}$ is larger. The third assumption is a 'mutual exclusion' prior. We observe that in categories that are hard to distinguish, e.g. `cat` and `dog`, `horse` and `cow`, a segment often has significant predicted overlaps on multiple categories, but only one of them is correct (see (Li et al., 2013) for an example). In such cases, when evidence from θ_{-j} suggests the object is from another category, the probability of $z_{ij} = 1$ is diminished by a factor (details in (Li et al., 2013)). The 1-dimensional function $\alpha(\hat{V}_{ij})$ is obtained by computing the histogram on the false positive rate over a validation set and fitting a smooth function to it.

3.3. EM Estimation

To maximize the likelihood with latent variable z_{ij}, we adopt a conventional expectation maximization (EM) approach. Formally, we would like to optimize the composite likelihood with latent variables $Z = [z_{ij}]$: In the E-step, $\mathbb{E}(z_{ij})$ is computed from existing estimates using Bayes' formula (see (Li et al., 2013)):

$$\mathbb{E}(z_{ij}) = p(z_{ij} = 1|\hat{V}, V, \theta) = f(V_{ij}, \theta_{-j}) p(z_{ij} = 1|\hat{V}_{ij}, V_{ij}) \tag{8}$$

This turns out to be similar to a standard mixture model update rule, with an additional factor $f(V_{ij}, \theta_{-j})$ reflecting the change of belief on V_{ij} and $\hat{V}_{ij}$, given current estimates of θ for other categories.

[1]Here $\mathcal{N}$ is a truncated Gaussian on the range $[0, 1]$, but since the log-likelihood between truncated and normal Gaussians differs only by a constant, we abuse the notation $\mathcal{N}$ here.

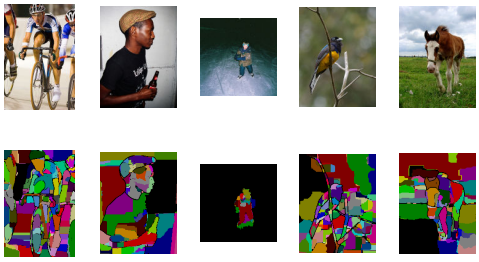

Figure 4. (Best viewed in color) Refined superpixels obtained by multiple intersection from original mid-level segments. Each different color represents a different superpixel (black identifies the largest one). Note that the partitions are, automatically, finer-grained, on the objects of interest.

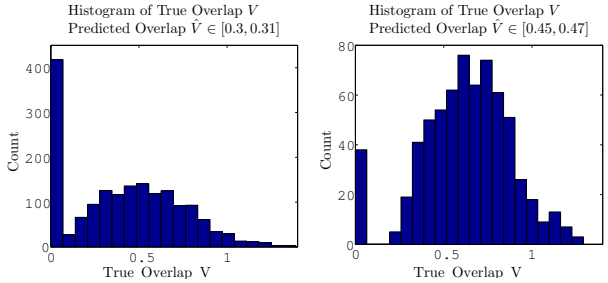

Figure 5. Histograms of true overlap given predicted overlap across the VOC validation set.

In the M-step we maximize the log-likelihood based on the following optimization (detailed derivations in (Li et al., 2013)):

$$\begin{aligned} \min_{\theta} \quad & \textstyle\sum_{i,j} \beta_{ij} \left(\frac{\mathbb{E}(z_{ij})}{2\sigma^2} (\hat{V}_{ij} - V_{ij}(\theta))^2 + (1 - \mathbb{E}(z_{ij})) \lambda V_{ij}(\theta) \right) \\ \text{s.t.} \quad & 0 \leq \theta_{kj} \leq 1, k = 1, \ldots, n, j = 1, \ldots, C; \\ & \textstyle\sum_{j=1}^{C} \theta_{kj} \leq 1, k = 1, \ldots, n \end{aligned} \tag{9}$$

Substituting (4) into (9) results in the full optimization. In practice, we also employ a smoothness regularization term $\lambda_2 \sum_{k=1}^{n} |S_k| \left(\sum_{j=1}^{c} \theta_{kj}^2 \right)$ where λ_2 is a parameter. It tends to preserve the shape of segments in the superpixel potentials and proved important for practical performance.

$$\max_{\theta, Z} \sum_{i=1}^{m} \sum_{j=1}^{r} \beta_{ij} \log p(\hat{V}_{ij}|V_{ij}(\theta), z_{ij}) \tag{10}$$

Interestingly, the optimization has a convex relaxation (see (Li et al., 2013)). In the M-step of each EM iteration we first solve the convex relaxation, then use the solution to warm start the optimization (9). A projected quasi-Newton method [2] is used to solve both optimization problems.

[2]http://www.di.ens.fr/ mschmidt/Software/minConf.html

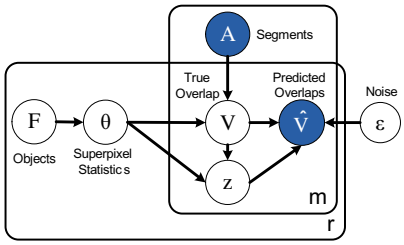

Figure 6. The graphical model used. We separate objects within each category ((Li et al., 2013)) so that the categorical predictions are mapped to each object. Also, θ and V generate a Bernoulli random variable z, which determines whether the predicted overlap would be a false positive.

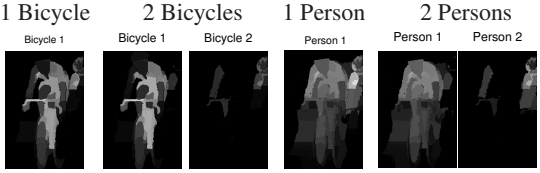

Figure 7. The θ map computed for 1 bicycle/2 bicycles, and 1 person/2 persons for the same set of predicted segment overlaps. The second bike represents spurious predictions from noise, whereas separating two people indeed improves the solution.

3.4. The Full Procedure

The full inference procedure involves two steps:

- Determining the number of objects within each category by the within-class object separation routine (see (Li et al., 2013)).
- Performing joint inference by iterating (8) and (9) across all categories and objects.

Notice that we choose to perform the within-class object separation routine before the joint inference, because enumeration of object counts in each category in a joint inference could lead to exponential blowups. Whereas, even if the within-class object separation can make mistakes, the erroneous object hypotheses can still be suppressed during the joint inference.

In fig. 7 we show the result of the within-class object separation routine on the segments in fig. 1. Two objects are generated in both the bicycle and the person categories. After detecting two objects for each category and running joint inference with these 4 objects, the algorithm is able to correct the mistake in the second bike (fig. 8).

4. Optimal Full Image Labeling

Given the inferred real-valued parameters θ (e.g. fig. 8), we still need to produce a consistent segment for each object. A graph-cut algorithm can be used on a potential map like fig. 8, but because θ has different magnitudes in different images, a uniform cut parameter choice across a dataset is unlikely to be successful. We propose an algorithm to produce optimal segments that maximizes the overlap with ground truth, without the need to re-segment. First, we prove that if there is only a single object, then the following procedure is optimal (see (Li et al., 2013) for details):

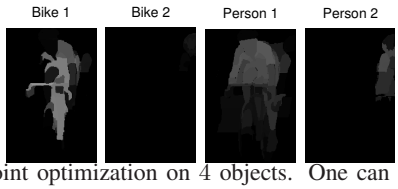

Figure 8. Joint optimization on 4 objects. One can see that potentials for Bicycle 2 have been suppressed due to similar spatial layout and lower scores to Person 2.

Figure 9. Final masks and final output of the algorithm. Bike 2 is filtered out because of very low score. Superpixels are chosen according to the procedure in Sec. 4. It is interesting to see that the first person has his right leg correctly cut through by the bicycle, a solution that was not available in any of the initial object segmentation proposals.

- Sort all θ in descending order. Initialize $C = 0$.
- From the start of the sorted list, include superpixel into the final segment one-by-one and compute the overlap V of the current segment using formula (4) from θ.
- Stop when $V > \frac{\theta_j}{1-\theta_j}$, and output the segment with superpixels 1 to $j-1$ in the sorted list.

In case the optimal segments in multiple categories conflict on some superpixels, one can run a branch-and-bound search on all the conflicting superpixels to maximize the sum of overlaps on each object (see (Li et al., 2013). Since θ from all categories are optimized jointly, only a limited number of superpixels will be simultaneously present in the optimal segment of many categories and the search is usually fast. Fig. 9 shows the search results for the 4 objects in fig. 8 as well as the final output.

5. Experiments

The experiments are conducted on the PASCAL VOC Segmentation dataset (Everingham et al., 2012), a widely used benchmark for semantic segmentation. This dataset defines 20 object categories and provides around $3,000$ training images with pixelwise ground truth annotations. In addition, around $9,000$ images annotated with bounding box information are used for training. The final benchmark is on a held out test set, for which an evaluation can only be done by submitting results to an online evaluation server. Performance is evaluated as the average pixel precision, computed on all the pixels of each class and then averaged over the 20 classes plus background. The overlap predictions $\hat{V}$ used in our system are obtained by combining the regressors from (Li et al., 2010) and (Carreira et al., 2012), with linear weights learned on the `trainval` set.

Table 1. VOC 2012 test results. CSI performs better especially in the categories with more object interactions.

Method	Mean	Background	Airplane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining Table	Dog	Horse	Motorbike	Person	Potted Plant	Sheep	Sofa	Train	TV/Monitor
SVRSEGM	46.8	84.9	63.8	22.1	50.5	38.9	44.8	61.3	63.3	48.8	9.8	57.2	35.6	43.0	51.1	58.8	53.7	29.7	49.8	30.3	47.0	38.0
JSL	47.0	85.1	65.4	29.3	51.3	33.4	44.2	59.8	60.3	52.5	13.6	53.6	32.6	40.3	57.6	57.3	49.0	33.5	53.5	29.2	47.6	37.6
CSI	47.5	85.2	64.0	32.2	45.9	34.7	46.3	59.5	61.6	49.4	14.8	47.9	31.2	42.5	51.3	58.8	54.6	34.9	54.6	34.7	50.6	42.2

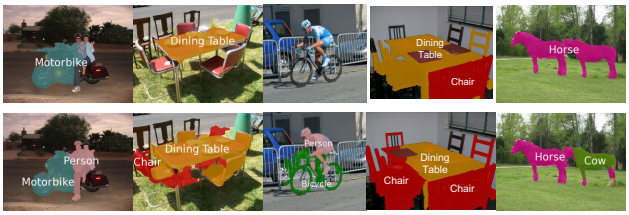

Figure 10. Example of semantic segmentations. The first row shows results using the post-processing algorithm of (Li et al., 2010), the second row shows results of the proposed CSI algorithm. Note CSI performs significantly better in complex scenes with multiple interacting objects.

In the results (Table 1) The method performs slightly better than the others, especially for object categories involved in interactions such as `Bike`, `Chair`, `Person` and `Sofa`. The 47.5% overall result for CSI is the best reported on *comp5* of the VOC 2012 challenge so far (Everingham et al., 2012). We show some images on the VOC test set in fig. 10. It can be seen that CSI handles object interactions very well in many cases. More images and comparisons are given in our technical report(Li et al., 2013).

6. Conclusion

This paper proposes a composite statistical inference approach to semantic segmentation. The composite likelihood methodology is generalized to model one-dimensional error distributions of statistical estimates. Then, superpixel-level inference is performed using a set of mutually overlapping object segmentation proposals and their predicted overlaps with object categories. The generative process is modeled and an EM algorithm is proposed to maximize the composite likelihood in two steps: the number of objects in each category is first determined, then a joint optimization is performed for all objects across categories. Once superpixel-level parameters have been estimated, the optimal pixel-level segmentation can be computed efficiently by branch-and-bound search. Experiments demonstrate the effectiveness of the approach, especially in scenes with multiple objects and interactions.

References

Arbelaez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L., and Malik, J. Semantic segmentation using regions and parts. In *CVPR*, 2012.

Carreira, J. and Sminchisescu, C. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *PAMI*, 2012.

Carreira, Joao, Caseiro, Rui, Batista, Jorge, and Sminchisescu, Cristian. Semantic segmentation with second-order pooling. In *ECCV*, 2012.

Dann, Christoph, Gehler, Peter V., Roth, Stefan, and Nowozin, Sebastian. Pottics the potts topic model for semantic image segmentation. In *DAGM*, 2012.

Dillon, Joshua V. and Lebanon, Guy. Stochastic composite likelihood. *J. Mach. Learn. Res.*, pp. 2597–2633, 2010.

Everingham, M., Van Gool, L., Williams, Chris, Winn, J., and Zisserman, A. The pascal visual object classes challenge 2012. www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

Gould, Stephen, Gao, Tianshi, and Koller, Daphne. Region-based segmentation and object detection. In *NIPS*, 2009.

Hoiem, Derek, Efros, Alexei A., and Hebert, Martial. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007.

Ion, Adrian, Carreira, João, and Sminchisescu, Cristian. Probabilistic join segmentation and labeling. In *NIPS*, 2011.

Kumar, M. P., Turki, H., Preston, D., and Koller, D. Learning specific-class segmentation from diverse data. In *ICCV*, 2011.

Ladicky, L., Russell, C., Kohli, P., and Torr, P. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

Ladicky, Lubor, Sturgess, Paul, Alahari, Karteek, Russell, Chris, and Torr, Philip. What,where and how many? combining object detectors and crfs. In *ECCV*, 2010.

Li, Fuxin, Carreira, João, and Sminchisescu, Cristian. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010.

Li, Fuxin, Carreira, Joao, Lebanon, Guy, and Sminchisescu, Cristian. Composite statistical inference for semantic segmentation. Technical report, Georgia Institute of Technology, April 2013.

Lindsay, BG. Composite likelihood methods. *Contemporary Mathematics*, 1988.

Pantofaru, Caroline, Schmid, Cordelia, and Hebert, Martial. Object recognition by integrating multiple image segmentations. In *ECCV*, 2008.