Composite Statistical Inference for Semantic Segmentation

Fuxin Li[(1)], Joao Carreira[(2)], Guy Lebanon[(1)], Cristian Sminchisescu[(3)]
(1) Georgia Institute of Technology. (2) ISR - University of Coimbra. (3) Lund University
[fli,lebanon]@cc.gatech.edu, joaoluis@isr.uc.pt, cristian.sminchisescu@math.lth.se

Abstract

In this paper we present an inference procedure for the semantic segmentation of images. Different from many CRF approaches that rely on dependencies modeled with unary and pairwise pixel or superpixel potentials, our method is entirely based on estimates of the overlap between each of a set of mid-level object segmentation proposals and the objects present in the image. We 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 and refine *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 of complex object interactions.*

1. Introduction

The goal of semantic segmentation is 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. This task is of great practical importance, because determining object boundaries is crucial for scene understanding and robot vision. However, the level of detail required makes inference extremely challenging and has stimulated interesting research in recent years [1, 8, 9, 11, 12, 13, 15, 19, 21].

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.

The existence of predictions for many mutually overlapping segments poses a new inference challenge for pixel labeling. Standard inference approaches in a high-order (hierarchical) CRF model [14, 15] can model both pixel/superpixel and segment-level layers with pixel/superpixel nodes and segment nodes interconnected based on overlap and compositionality. However, the interactions in these models are complex and involve different types of pairwise potentials (between pixels, between pixels and segments and between segments) which limits the range of potential functions for which tractable approximate inference is feasible. A recently proposed variation using latent topics, the Pottics model [5], sidesteps the need for high-order cliques but still requires approximate inference.

Other approaches search for configurations of non-overlapping segment hypotheses [9, 13] by using non-maxima suppression and maximum clique random field models [11]. They can be tractable since the decision space that has to be searched is limited to the initial segments (normally < 200 in practice). However, these are likely to encounter difficulties when multiple objects touch or interact with each other. Examples are: people riding bicycles or horses, interacting with other people, sitting on sofas or chairs, etc. In such cases, segments often occlude and cut through each other and the initial mid-level proposals may not be entirely accurate. In such situations, a high-precision approach should be able to refine the initial hypotheses.

Non-probabilistic methods have also been developed to produce an average [10, 19] or weighted average [4, 17] of the predicted scores on each pixel/superpixel, then output the highest scoring labels. Arbelaez *et al*. learn to classify superpixels using class predictions from all enclosing segments as input features [1]. This strategy would typically allow for the refinement of a semantic segmentation 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 struc-

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.424

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

tures. A main deviation from CRF approaches is that instead of directly modeling the conditional 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 transfer conflicting high-order terms into a number of one-dimensional distributions, hence avoiding difficult maximum a posteriori inference in 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 intuitions obtained through 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 composites* of statistics on superpixels that do not spatially overlap. By computable, we mean there exists a mathematical formula that can output segment statistics given values of the superpixel statistics. Based on such a 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 of predicted segment statistics will emerge as the maximum likelihood solution, as shown in fig. 2. The generative graphical model is presented in fig. 3 and encodes the dependency of the ground truth statistic on the segments and the superpixel statistics, as well as the dependency of the observations on predicted segment statistics and a noise source.

Our methodology consists of a training phase and an inference phase. In the training phase, regressors are estimated to predict segment statistics. This can be done by standard routines such as SVR or least-squares, and is not

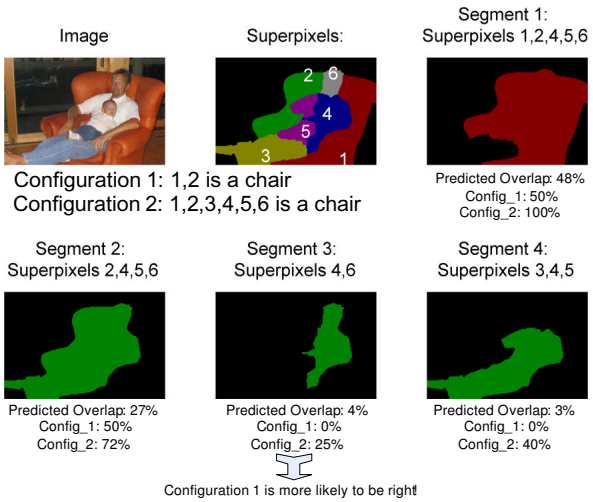

Figure 2. (Best viewed in color) The goal of our inference can be intuitively thought as finding the superpixel configuration which best explains most of the predicted segment statistics, here spatial overlap (with the chair object). This formulation allows discovering objects that are cut into disconnected components, such as the chair. Instead of find such a superpixel configuration using a search algorithm, we formulate it as a continuous maximum composite likelihood problem with a convex relaxation, where a near-optimal solution can be found via mathematical optimization.

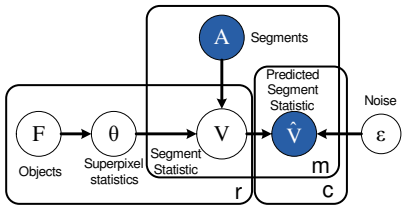

Figure 3. The conceptual graphical model. Superpixel statistics are generated from the ground truth objects. Segment statistics are generated from superpixel statistics and the segments. The observations are predicted segment statistics on each category. They are the maximal segment statistic for all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

covered here. Given a test image, the inference phase has three main stages:

- Use the trained 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. Since this is a new methodology, many innovations are presented in the paper to facilitate its execution in the difficult semantic segmentation problem:

- To ensure the computable composite assumption, our superpixels are obtained by multiple intersections from

the mid-level segment hypotheses, so that each superpixel either totally belongs to a segment or is completely outside it.

- We generalize the composite likelihood methodology to handle statistic estimates instead of probabilistic estimates.
- We introduce a prior on the number of objects. A maximum-a-posteriori (MAP) step is used to infer the optimal number of objects within each category.
- An EM algorithm is used to maximize the likelihood based on the mixture error model. The E-step assigns mixture weights and the M-step maximizes the composite likelihood. In the M-step, we propose a good convex relaxation which is used to warm-start the solution.
- For the last stage, we exploit the structure in the superpixel statistics in order to propose an efficient, optimal search algorithm to find the best pixel labeling given the estimated superpixel statistics.

2. Composite Likelihood and Its Generalization to Statistical Estimates

Throughout the paper we denote $p(x)$ the probability of random variable x, $\mathbb{I}$ the indicator function. $\mathcal{N}(x;\mu,\sigma^2)$ the density function of the normal distribution with mean μ and variance σ^2, $Ber(\alpha)$ a Bernoulli distribution with parameter α, $\text{Exp}(x;\lambda)$ the density of 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 [18, 20] 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. We adopt a version in [6] with some simplifications, and refer the reader to [6] and our associated technical report [16] for details.

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

$$cl(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{k} \beta_j \log p_\theta(X^{(i)}_{A_j} | X^{(i)}_{B_j}). \qquad (1)$$

When β has stochastic components, this is called stochastic composite likelihood (SCL)[6]. MCL is the approach to solve for θ by maximizing the composite likelihood (1). The MCL/SCL approach is statistically consistent given an identifiability assumption, but is intractable in most cases, however, 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. This makes us work with 1-dimensional distributions which are much 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*

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

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. It seeks to find a parameter vector θ that best explains all the observed statistics from $X^{(i)}$ and the two given subsets A_j and B_j. The distribution $p_\theta(f(X^{(i)}, A_j, B_j))$ is modeled as a 1-dimensional distribution. A relevant example is a linear subset regression model with Gaussian errors. Suppose $X^{(i)}$ is an image, A_j is a subset of its pixels (a segment) and B_j is a background segment non-overlapping with A_j. Then a fixed-length feature vector Z_{ij} can be extracted from these segments 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. Given observed values of f_{ij} for many different $X^{(i)}, A_j$ and B_j, the MCL problem in this case becomes a weighted least squares regression of solving for θ. As shown in our associated technical report[16], the asymptotic consistency proof still partially holds, 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 main parametric model of the proposed CSI (Composite Statistical Inference) method that uses the modified MCL to infer semantic segmentations. We will first present the probabilistic model (sec. 3.2), followed by the EM algorithm to estimate parameters (sec. 3.3). We must convert the per-category scores to per-object scores in order to properly maximize the likelihood. To do so, we need to estimate the number of objects in each category and assign the score of each segment belonging to

a particular object. We postpone the relevant discussion to sec. 3.4 because it uses the same probabilistic model and EM formulation introduced in sec. 3.2 and 3.3. A discussion on how to output final segmentations given the superpixel statistics estimated from MCL is deferred to 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

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

The true overlap V_{ik}^0 can be estimated by training one regressor for each category C_k(for details on possible training methods one can consult e.g. [17, 4, 1]). Since this paper deals with inference, which is only required during testing, we assume that regressors are already obtained based on a separate training set and denote their estimates in the test image I 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. 5) 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

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

Importantly, V_{ij} is computable from θ only since each $|S_k|$ is a constant. The idea is that if one parameterizes the ground truth object with θ, then its overlap with each segment can be computed (fig. 2). 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:

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

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. In order to simplify the presentation of the graphical model, we assume for now that this assignment has been resolved, so that 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:

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

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. Note that an assignment is performed within each category, hence a segment can be assigned to many objects, but at most 1 per category. The resolution of the assignment problem within each category will be described in Sec. 3.3.

We assume that the estimated overlap $\hat{V}_{ik}$ is generated from the true overlap V_{ik} plus noise. In order to determine the form of $p(\hat{V}|V)$, we resort to histograms. Fig. 6 shows histograms on $V|\hat{V}$, for the data collected from PASCAL VOC training set. The distribution of $V|\hat{V}$ 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.

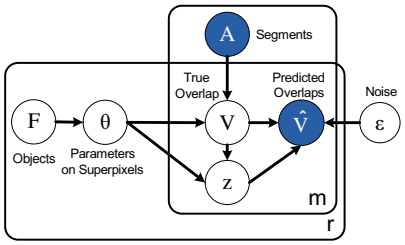

Figure 4. The graphical model used. We separate objects within each category (Sec. 3.4) so that the categorical predictions are mapped to each object. Also, θ and V generate a Bernoulli random variable z, which determines whether the predicted overlap would be a false positive.

Motivated by these observations, we introduce an additional Bernoulli random variable z_{ij} for each predicted score $\hat{V}_{ij}$ (fig. 4). 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 $\theta_{-j} = [\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_r]$ represents all the θ columns without the j-th. These assumptions are in

line with our observations: if $z_{ij} = 0$, the prediction is a false positive and the true 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 our technical report [16] for an example). In such cases, when we have evidence from θ_{-j} that an object in another category might exist, the probability of $z_{ij} = 1$ is diminished by a factor (details in [16]). 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.

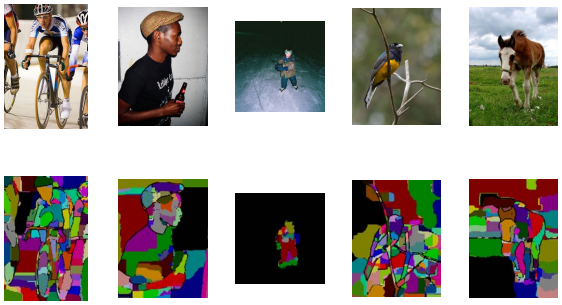

Figure 5. (Best viewed in color) Refined superpixels obtained by multiple intersection from original mid-level segments. Each different color represents a different superpixel (black identifies the largest one). Note that the partitions are, automatically, finer-grained, on the objects of interest.

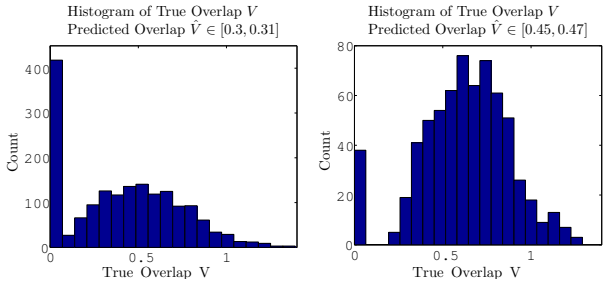

Figure 6. Histograms of true overlap given predicted overlap across the VOC validation set. One can easily identify two components: a probability mass at 0 and a centered distribution to the right. The 0 mass corresponds to misclassifications, where the object does not belong to the category, but the regressor erroneously outputs nonzero predicted overlaps. Also note that with higher predicted overlap $\hat{V}$, there is less chance for $V = 0$.

If the prediction is biased such that $\mathbb{E}(V|\hat{V}) \neq \mathbb{E}(\hat{V})$, we could correct the (systematic) bias in this step by using a function $g(\hat{V}_{ij})$ instead of $\hat{V}_{ij}$ in the assumption. The bias-correction function could be fitted in the same way as α, by taking a histogram on the validation data and smoothing it. Because both α and g are 1-dimensional functions, the risk of overfitting is drastically reduced when the validation set becomes large (*e.g.* with cross-validation).

[1]Here $\mathcal{N}$ should be viewed as 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.

3.3. EM Estimation

To maximize the likelihood with latent variable z_{ij}, we adopt a conventional expectation maximization (EM) approach. In the E-step, we will average over choices of z_{ij}, and then in the M-step maximize the expected log-likelihood. Formally, we would like to optimize the composite likelihood with latent variables $Z = [z_{ij}]$:

$$\max_{\theta, Z} \sum_{i=1}^{m} \sum_{j=1}^{r} \beta_{ij} \log p(\hat{V}_{ij}|V_{ij}(\theta), z_{ij}) \tag{8}$$

In the E-step, $\mathbb{E}(z_{ij})$ is computed from existing estimates using Bayes' formula (see [16]):

$$\mathbb{E}(z_{ij}) = p(z_{ij} = 1|\hat{V}, V, \theta) = f(V_{ij}, \theta_{-j}) p(z_{ij} = 1|\hat{V}_{ij}, V_{ij}) \tag{9}$$

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.

In the M-step we maximize the log-likelihood based on the following optimization (detailed derivations in [16]):

$$\begin{aligned} \min_{\theta} \quad & \textstyle\sum_{i,j} \beta_{ij} \left(\frac{\mathbb{E}(z_{ij})}{2\sigma^2} (\hat{V}_{ij} - V_{ij}(\theta))^2 + (1 - \mathbb{E}(z_{ij})) \lambda V_{ij}(\theta) \right) \\ \text{s.t.} \quad & 0 \leq \theta_{kj} \leq 1, k = 1, \ldots, n, j = 1, \ldots, C; \\ & \textstyle\sum_{j=1}^{C} \theta_{kj} \leq 1, k = 1, \ldots, n \end{aligned} \tag{10}$$

Substituting (4) into (10) results in the full optimization. Since θ_{kj} models percentages, it has a range of $[0, 1]$, represented in the first constraint. The second constraint comes from the assumption that each pixel can belong to only 1 object. In practice, we also employ a 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. This regularizer can be viewed as a smoothness term that promotes a more uniform selection of θ. It tends to preserve the shape of segments in the superpixel potentials and proved important for practical performance.

Interestingly, the optimization has a convex relaxation. The expanded form of both (10) and its convex relaxation are given in our associated report[16]. In the M-step of each EM iteration we first solve the convex relaxation, then use the solution to warm start the optimization (10). A projected quasi-Newton method from `minConf`[2] is used to solve both optimization problems.

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

3.4. Locating Multiple Objects within Each Category

To locate multiple objects in one category and in order to separate the estimates to each object, we adopt the above EM estimation with a hypothesis-testing framework to find the number of objects in each category, in a MAP setting. Namely, we solve (2) for each category C_k independently, with an additional geometric prior on the number of objects r_k: $p(r_k = j) = (1-q)^j q$, where $q > 0$ is a parameter. For each of $r_k = 1, 2, 3$, etc., the following posterior is computed:

$$L_{r_k} = \max_{\theta, Z} \sum_{i=1}^{m} \max_{j \in 1,\ldots,r} \log p(\hat{V}_{ik}^0 | V_{ij}(\theta), z_{ij}) + r_k(1-q) \quad (11)$$

by maximizing over θ and Z. The posteriors L_{r_k} are computed iteratively. First L_1 is computed by setting all $\mathbb{E}(z_{i1}) = \alpha_{ik}$ and running the M-step (10) only. Then, suppose L_{r_k} is computed with the optimized parameters as θ_{r_k}, Z_{r_k}, L_{r_k+1} is inductively computed by adding one object with an initialization of:

$$\mathbb{E}(z_{i,r_k+1}) = 1 - \frac{\max_{j \in \{1,\ldots,r_k\}} p(\hat{V}_{ik}^0 | V_{ij}(\theta_r))}{p(\hat{V}_{ik}^0 | V = \hat{V}_{ik}^0)} \quad (12)$$

and running the EM steps (10) and (9) until convergence. In (12), the denominator represents the maximum likelihood from any configuration, and the nominator represents the likelihood of the best explanation of $\hat{V}_{ik}^0$ by any of the current j objects. The logic behind (12) is that, if $\hat{V}_{ik}^0$ has already been explained perfectly, adding an object cannot improve the likelihood thus $\mathbb{E}(z_{i,r_k+1})$ is initialized to 0. If none of the objects has been able to explain $\hat{V}_{ik}^0$ so far, then a new object is likely present, thus $\mathbb{E}(z_{i,r_k+1})$ is initialized to 1.

At any point, if $L_{r_k+1} < L_{r_k}$, the computation is stopped and r_k is decided to be the number of objects. Then, each segment is assigned to the object F_j that maximizes $\mathbb{E}(z_{i,j})$ in the final Z_{r_k}. The joint inference on all categories is subsequently performed, by treating each object as a different category with separately assigned predictions.

3.5. 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 in Sec. 3.4.
- Performing joint inference by iterating (9) and (10) across all categories and objects.

Notice that we choose to perform the within-class object separation routine before the joint inference, because within each category the enumeration of object counts is tractable. If one enumerates in the joint inference phase, then hypotheses like "1 object in c_1, 2 objects in c_2" need to be tested and could lead to exponential blowup when there are many categories. Whereas, even if the within-class object separation can make mistakes, the erroneous object hypotheses can still be suppressed during the joint inference.

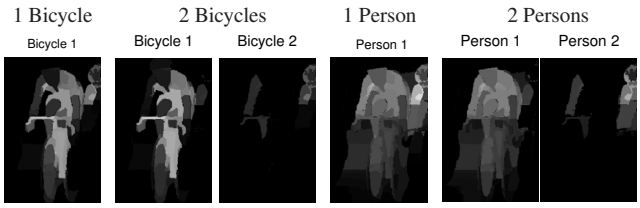

Figure 7. Different θ computed for 1 bicycle/2 bicycles, and 1 person/2 persons hypotheses for the same set of predicted segment overlaps. The second bike represents spurious predictions from noise, whereas separating two people indeed improves the solution.

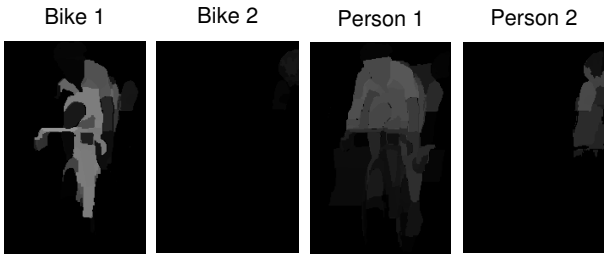

Figure 8. Joint optimization on 4 objects. One can see that potentials for Bicycle 2 have been suppressed due to similar spatial layout and lower scores to Person 2.

In fig. 7 we show the result of running the within-class object separation routine on the segments in fig. 1. One can see that in both the bicycle and the person categories, two objects are generated instead of one. Although both categories improve the likelihood by predicting 2 objects, the second bicycle object is erroneous whereas the second person object is correct. After detecting two objects for each category and running joint inference with these 4 objects, the algorithm is able to correct that mistake, as shown in 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, note that the overlap formula (4) can also be written as:

$$V(F_j, A) = \frac{\sum_{S_k \in A} \theta_{kj} |S_k|}{\sum_{k=1}^{n} \theta_{kj} |S_k| + \sum_{S_k \in A} (1 - \theta_{kj}) |S_k|} \quad (13)$$

where in the denominator we first count all the ground truth pixels in F_j by $\sum_{k=1}^{n} \theta_{kj} |S_k|$, then sum all the pixels inside segment A_i that do not belong to F_j. This reformulation

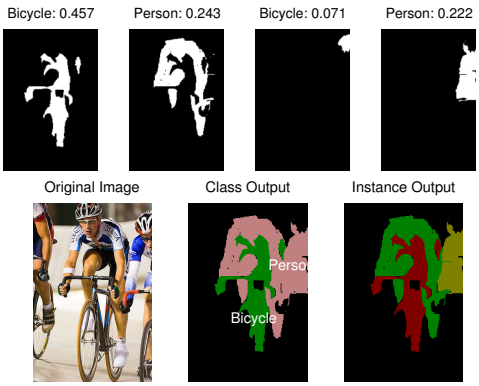

Figure 9. Final masks and final output of the algorithm. Bike 2 is filtered out because of very low score. Not all superpixels with non-zero potentials are in the final mask, because adding some more would be suboptimal according to the procedure in Sec. 4. It is interesting to see that the first person has his right leg correctly cut through by the bicycle, a solution that was not available in any of the initial object segmentation proposals.

leads to a simple approach to grow A optimally. Suppose we have A with $V(F_j, A) = V_0$, then V can be increased if and only if we add a superpixel to A with $\frac{\theta_{kj}}{1-\theta_{kj}} > V$, because $\frac{a+c}{b+d} > \frac{a}{b}$ iff $\frac{c}{d} > \frac{a}{b}$. Therefore, when the image contains only an object in a single category, the optimal segment can be found by starting from $A = \emptyset$ and $V = 0$. We then sort $\frac{\theta_{kj}}{1-\theta_{kj}}$ corresponding to all superpixels in descending order, and keep adding superpixels from the top of the list until $V \geq \frac{\theta_{kj}}{1-\theta_{kj}}$ for all remaining superpixels.

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. For each conflicting superpixel S_k, a quality function is defined by

$$Q_{kj} = \max_{A} V(F_j, A) - \max_{A, S_k \notin A} V(F_j, A) \qquad (14)$$

where we perform the search in a best-first manner, with the superpixel S_k for object F_j picked first if the pair has the best quality Q_{kj}. At each branch, an upper bound is computed by $\max_{A, S_k \subset A} V(F_j, A)$ and a lower bound is computed by $\max_{A, S_k \cap A = \emptyset} V(F_j, A)$, where max can choose from all the superpixels that have not been assigned at the branch. These bounds prune the search space effectively.

The search can be performed very fast because: 1) Since θ from all categories are optimized jointly, one superpixel is likely to be assigned to a single category and only a limited number of superpixels will be simultaneously present in the optimal segment of many categories; 2) The bounds obtained with the above procedure are usually quite tight. In many cases, a greedy approach using the quality function achieves the optimal solution. 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 [7], 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. This set, named `trainval`, was further divided into half in the `train` and half in the `val` set. In addition, around $9,000$ images annotated with bounding box information can be used for training. The final benchmark of performance is a held out `test` set, for which the ground truth is not available and 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. We tune the parameters λ, λ_2 and δ and the α function on the `val` set using the regressor output trained on `train` and the additional images with bounding box annotations. Then, evaluation is performed on the `test` set with the tuned parameters and fitted functions. The overlap predictions $\hat{V}$ used in our system are obtained by combining the regressors from [17] and [2], with linear weights learned on the `trainval` set. The parameters λ, λ_2 and δ are tuned on the `val` set.

On the VOC `test` set, we compare the proposed CSI approach against other methods on the 2012 challenge using the same set of category prediction scores, which includes SVRSEGM [3] and JSL [11]. The JSL entry to VOC 2012 is different from the paper [11] in that it also employed pixel-level averaging to improve performance. It can be seen from Table 1 that the method performs slightly better than the others, especially for object categories involved in interactions such as `Bike`, `Chair`, `Person` and `Sofa`. It does less well in the animal categories where interactions are less likely to happen. The 47.5% overall result for CSI is the best reported on *comp5* of the VOC 2012 challenge so far [7].

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[16].

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. Based on this generalization, superpixel-level inference is performed based on a set of mutually overlapping object segmentation proposals and their predicted overlaps with object categories. The generative process underlying overlap prediction is modeled using a graphical model

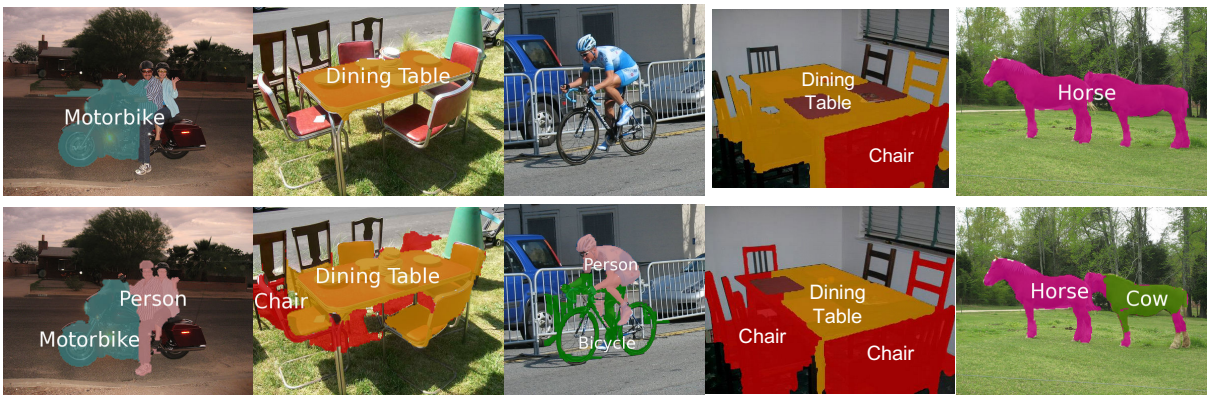

Figure 10. Example of semantic segmentations. The first row shows results using the post-processing algorithm of [17], the second row shows results of the proposed CSI algorithm. Areas of the image labeled as background are depicted with their original appearance. The first four images show cases where our algorithm is more accurate, mainly involving relatively complex scenes with multiple interacting objects. The last image, on the right, shows a typical failure case: segments covering part of one of the horses are strongly confused and assigned to 'cow'. The algorithm of [17] typically oversmooths the predictions, which is advantageous in some cases, like in this image.

Table 1. VOC 2012 test results

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

and an EM algorithm is proposed to solve the maximum composite likelihood inference 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 best-first search. Experiments demonstrate the effectiveness of the approach, especially in scenes with multiple objects and interactions.

Acknowledgements: The authors thank Joshua Dillon for helpful discussions. This work was supported in part by NSF project IIS-1016772 and FCT project PTDC/EEA-CRO/122812/2010.

References

[1] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, 2012. 1, 4

[2] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 7

[3] J. Carreira, F. Li, and C. Sminchisescu. Object Recognition by Sequential Figure-Ground Ranking. *IJCV*, 2012. 7

[4] J. Carreira and C. Sminchisescu. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *PAMI*, 2012. 1, 4

[5] C. Dann, P. V. Gehler, S. Roth, and S. Nowozin. Pottics the potts topic model for semantic image segmentation. In *DAGM*, 2012. 1

[6] J. V. Dillon and G. Lebanon. Stochastic composite likelihood. *J. Mach. Learn. Res.*, pages 2597–2633, 2010. 3

[7] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012. www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 7

[8] J. Gonfaus, X. Boix, J. V. de Weijer, A. Bagdanov, J. Serrat, and J. Gonzlez. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010. 1

[9] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*, 2009. 1

[10] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007. 1

[11] A. Ion, J. Carreira, and C. Sminchisescu. Probabilistic join segmentation and labeling. In *NIPS*, 2011. 1, 7

[12] P. Krahenbuhl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 1

[13] M. P. Kumar, H. Turki, D. Preston, and D. Koller. Learning specific-class segmentation from diverse data. In *ICCV*, 2011. 1

[14] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 1

[15] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. Torr. What,where and how many? combining object detectors and crfs. In *ECCV*, 2010. 1

[16] F. Li, J. Carreira, G. Lebanon, and C. Sminchisescu. Composite statistical inference for semantic segmentation. Technical report, Georgia Institute of Technology, April 2013. 3, 5, 7

[17] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010. 1, 4, 7, 8

[18] B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 1988. 3

[19] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations. In *ECCV*, 2008. 1

[20] C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92, 2005. 3

[21] W. Xia, Z. Song, J. Feng, L.-F. Cheong, and S. Yan. Segmentation over detection by coupled global and local sparse representations. In *ECCV*, 2012. 1