Object Co-detection via Efficient Inference in a Fully-Connected CRF*

Zeeshan Hayder, Mathieu Salzmann, and Xuming He

Australian National University (ANU)
NICTA, Canberra, Australia

Abstract. Object detection has seen a surge of interest in recent years, which has lead to increasingly effective techniques. These techniques, however, still mostly perform detection based on local evidence in the input image. While some progress has been made towards exploiting scene context, the resulting methods typically only consider a single image at a time. Intuitively, however, the information contained jointly in multiple images should help overcoming phenomena such as occlusion and poor resolution. In this paper, we address the co-detection problem that aims to leverage this collective power to achieve object detection simultaneously in all the images of a set. To this end, we formulate object co-detection as inference in a fully-connected CRF whose edges model the similarity between object candidates. We then learn a similarity function that allows us to efficiently perform inference in this fully-connected graph, even in the presence of many object candidates. This is in contrast with existing co-detection techniques that rely on exhaustive or greedy search, and thus do not scale well. Our experiments demonstrate the benefits of our approach on several co-detection datasets.

Keywords: Object co-detection, fully-connected CRFs.

1 Introduction

Object detection has been a central problem in modern computer vision, and much progress has been made in recent years, as demonstrated by the PASCAL challenge [12] and the ImageNet challenge [7]. Whether working at instance [22] or category [14] level, most of the research has focused on detecting objects in a single image and in a sliding window manner. It is widely acknowledged, however, that such a myopic view is too restrictive as it ignores all contextual information [15]. On their own, the appearance cues of an object instance are often ambiguous due to poor resolution, occlusions, or challenging lighting conditions.

Previous work on object detection with context mainly exploits the 2D or 3D scene context observed in the same image as the detected objects [19,8]. Recently, simultaneously exploiting multiple images has been proposed as a means

* NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy, as well as by the Australian Research Council through the ICT Centre of Excellence program.

D. Fleet et al. (Eds.): ECCV 2014, Part III, LNCS 8691, pp. 330–345, 2014.

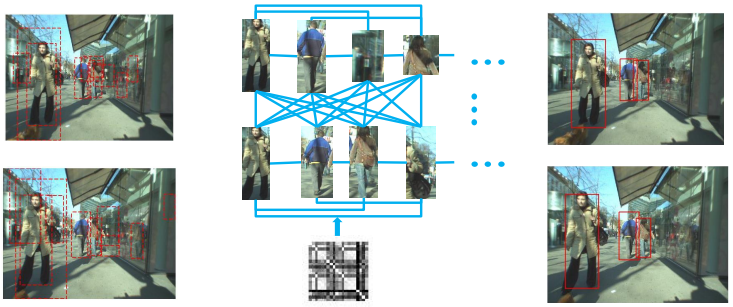

Fig. 1. Overview of our method. Left: Original input images and candidates generated with a DPM; Middle: Fully-connected CRF on the candidates and corresponding learned pairwise similarities; Right: Jointly detected objects by efficient inference in the fully-connected CRF (actual result).

to gather broader contextual information for detection. The resulting *object co-detection* techniques [3,18,24] aim to jointly detect multiple instances of an object class from a pool of images. Intuitively, object co-detection leverages the weak appearance cues of object instances seen in multiple images to improve the robustness of object detection.

A critical challenge in object co-detection is to incorporate many object hypotheses from multiple images while keeping the joint classification of those object hypotheses tractable. Typically, the problem is formulated as that of inferring the (binary) activation labels of object candidates, which is a combinatorial search problem. The existing methods rely on either exhaustive search [3], or ad hoc greedy search [24]. While these strategies are effective for a small number of images, they are in general suboptimal, and become impractical when considering large image pools or number of classes.

In this paper, we introduce a principled and efficient inference method for object co-detection. Given a pool of object candidates obtained by applying a pre-trained detector with a high recall rate (e.g., the Deformable Part-based Model (DPM) [14]), we construct a fully-connected Conditional Random Field (CRF) where the nodes represent the candidate labels, and the edges encode the appearance similarity between two candidates. Inference in this CRF lets us predict the labels of all the object candidates simultaneously.

For our formulation to remain tractable, we need to be able to leverage efficient inference techniques in fully-connected CRFs. To this end, we model the similarity between two candidates as a linear combination of Gaussian kernels defined on multiple image features. The weights of this combination can be efficiently learned from training data. We make use of this similarity in the edge potentials of our CRF, which encode a data-dependent Potts model. The form of these potentials lets us utilize the efficient mean field inference algorithm of [21], which not only yields the candidate labels, but also confidence in our predictions. Fig. 1 depicts an overview of our framework.

We evaluate our method on three benchmark co-detection datasets: the Pedestrian dataset [11], the Ford Car dataset [3] and the Human Co-Detection dataset [24]. In all three cases, our approach outperforms the state-of-the-art co-detection methods, thus demonstrating the benefits of adequately modeling the relations between all object candidates via our fully-connected CRF formulation.

2 Related Work

Object modeling and recognition has been one of the fundamental problems in computer vision since its early days [9]. In particular, object detection has evolved into a core challenge in vision research [12,7], and much progress has been made recently due to advances in deformable part-based object models, e.g., DPM [14], as well as in deep network models [16].

Traditionally, object detection methods take a scanning window approach and exhaustively search for object candidates at every location and scale in an image [28,6,14]. More recently, objectness criteria have been used to propose potential candidates, which drastically reduces the search space [10,1,25]. Objectness, however, is very challenging to adequately represent, since it has to account for the huge intra-class variations of the general *object* category, while still being able to differentiate it from the *background* class. As a matter of fact, these challenges remain unsolved even when detecting specific objects.

A natural perspective to improve the robustness of detectors to phenomena such as poor resolution, occlusions, and challenging lighting conditions, consists in putting objects into context. To this end, the scene properties of the target image are exploited to boost the object detection performance. For instance, Desai et al. [8] propose to jointly detect multiple object classes by defining a CRF on top of DPMs. In [4], multiple instances of the same object class are jointly detected to address the occlusion problem. Hoiem et al. [19] consider the geometric context of the scene to improve detection by reducing the number of false positives. While these approaches have proven more effective than context-free detectors, they focus on exploiting the context from a single image. Intuitively, however, the information available in multiple images should be helpful to disambiguate detection.

Object co-detection methods [3,18,24] were recently introduced to exploit the collective power of a set of images. In particular, the term *co-detection* was coined by Bao et al. [3], who tackle the problem by exhaustively searching for matching object instances in a set of object candidates. Generally speaking, co-detection has been considered for both 2D and 3D object models, as well as at category- and instance-levels. Category-level co-detection involves matching objects belonging to the same class (e.g., a person with another person) and appearing either in the same image, or in multiple images. In contrast, instance-level co-detection compares specific object instances (e.g., a specific person) that appear simultaneously in a pool of input images (e.g., [24]). While the original work of Bao et al. [3] could only handle pairs of images, Guo et al. [18] introduced a robust approach to multi-image co-detection that builds a shared low-rank representation of the object instances in multiple feature spaces. Unlike these works,

our method is based on a principled CRF formulation. Therefore, it enables us to perform joint inference efficiently for many object instances extracted from multiple images.

Our work is inspired by the fully-connected CRF model for semantic labeling of [21,27,30]. By restricting the functional form of the pairwise potentials to a weighted mixture of Gaussian kernels defined on the input feature space, inference in this fully-connected CRF can be performed efficiently as a filtering operation. Here, instead of labeling the pixels in an image, we aim to label object candidates from multiple images in a principled and yet efficient manner. To the best of our knowledge, our work is the first attempt to extend the fully-connected CRF model of [21] to another vision problem domain.

In this context, we propose to learn the pairwise potentials in our CRF by fitting a linear combination of kernels to a target similarity measure. This bears some connections with the multiple kernel learning literature [17,26]. However, our objective is not to build a kernel-based similarity classifier as in [24], since this would not yield a mixture of Gaussian kernels adapted to our fully-connected CRF framework. Instead, our similarity learning approach is closer to metric learning [29]. In contrast, however, we jointly consider multiple kernels defined on separate feature spaces, thus yielding more flexibility than the single linear transformation typically used in metric learning methods.

3 A Fully-Connected CRF for Co-detection

In this paper, we tackle the object co-detection problem, in which we aim to detect simultaneously all the instances of an object class in a group of S input images $\mathcal{I} = \{I^1, \cdots, I^S\}$. As in [3,18,24], when dealing with $C > 1$ object classes, we handle each class separately. Note that, while we discuss the case of category-level detection, our framework also applies to detecting the instances of specific objects (instance-level). Furthermore, we do not assume that each image contains only a single instance of an object class.

For each object class, our approach consists of two stages: first we generate a pool of object candidates in the form of bounding boxes obtained with a pre-trained object detector; then we formulate co-detection as a two-class labeling problem, where each candidate must be assigned either to the current object class of interest, or to a background class. These two steps are described in more detail in the remainder of this setion.

3.1 Object Candidate Generation

Following [3], given a target object class c, we first apply a pre-trained DPM [14] to each input image and extract a set of object candidates, denoted by $\mathcal{X}^c = \{\mathbf{X}_1^c, \cdots, \mathbf{X}_{N_c}^c\}$. To prevent entirely missing some objects in this first stage, we adjust the threshold of each detector and the non-maximum suppression parameters so as to achieve a high recall rate for all target classes. Note that, while we employ DPMs, any object detector that outputs a bounding box can

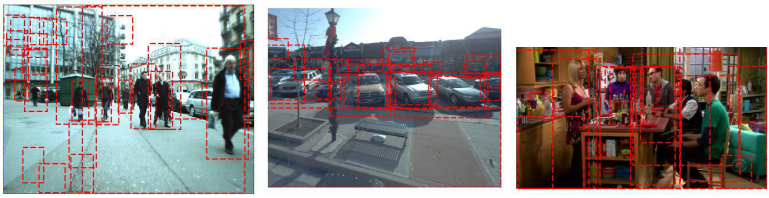

Fig. 2. Sample object candidates in three datasets. Left: Pedestrian dataset [11]; Middle: Ford Car dataset [3]; Right: Human Co-detection dataset [24].

be employed in our candidate generation stage. Fig. 2 shows some examples of object candidates generated for three different object classes.

Given the set of candidates $\mathcal{X}^c$ for class c, we then adopt a part-based representation as in the DPM. An object candidate $\mathbf{X}_i^c$ is represented by its root $\mathbf{r}_i^c$ and a set of k parts $\mathcal{P}_i^c = \{\mathbf{p}_{i,1}^c, \cdots, \mathbf{p}_{i,k}^c\}$, together with the image window $\mathbf{W}_i^c$ corresponding to the object bounding box. For each candidate $\mathbf{X}_i^c$, we also compute a set of appearance features from its image window $\mathbf{W}_i^c$. These features, denoted by $\mathbf{f}_{i,s}^c$ for a specific feature type s, capture the color and texture properties of the candidate.

3.2 CRF Formulation

Given the candidate pool $\mathcal{X}^c$, we formulate object co-detection as the problem of jointly labeling the candidates with the corresponding object or background class. More specifically, we introduce a label variable y_i^c for each object candidate $\mathbf{X}_i^c$, which takes either the object class label l^c, or the background label l^0.

To appropriately capture the dependencies of our object candidates, we build a fully-connected Conditional Random Field (CRF) on the label variables $\mathcal{Y}^c = \{y_1^c, \cdots, y_{N_c}^c\}$. Each node in the CRF corresponds to the label of one object candidate, and any pair of two candidates are connected by an edge that encodes their relationship. Formally, we define the joint distribution over the label variables $\mathcal{Y}^c$ given the observed candidates $\mathcal{X}^c$ as

$$P(\mathcal{Y}^c|\mathcal{X}^c) = \frac{1}{Z(\mathcal{X}^c)} \exp\left(-\sum_{i=1}^{N_c} \phi_u(y_i^c|\mathbf{X}_i^c) - \alpha \sum_{i=1}^{N_c} \sum_{j>i} \psi_p(y_i^c, y_j^c|\mathbf{X}_i^c, \mathbf{X}_j^c) \right), \quad (1)$$

where $Z(\cdot)$ is the partition function, α is a weight learned by cross-validation, and ϕ_u and ψ_p are the unary and pairwise potential functions, respectively. The unary potential ϕ_u encodes how likely a candidate is to be associated with each class, while the pairwise potential ψ_p measures the affinity between the different possible class assignments of two candidates.

Object co-detection then boils down to inferring the optimal label configuration of this CRF model, which jointly labels all the object candidates. In our work, we do not put any restriction on the number of input images. Consequently,

we may have a large number of object candidates (nodes) in our CRF. Inference in such a large, fully-connected CRF is in general intractable and difficult to approximate. The key challenge therefore lies in finding an efficient inference procedure in our fully-connected CRF.

To address joint inference in a principled way, we rely on the formulation of [21] to design our CRF model. The main requirement of this formulation is that the pairwise potentials must have the form of a mixture of Gaussian kernels. In the following, we discuss the potential functions employed in our model, and, in particular, introduce pairwise potentials that meet this mixture-of-Gaussian-kernels requirement and, as we will show, are effective for co-detection.

Unary Potentials. The unary potentials measure the likelihood that a candidate $\mathbf{X}_i^c$ belongs to the object class and to the background class. Following [3], we use a rescaled DPM score as unary potential. This lets us write our unary term as

$$\phi_u(y_i^c|\mathbf{X}_i^c) = \begin{cases} E_r(\mathbf{r}_i^c, \mathbf{W}_i^c) + \sum_{j=1}^{k} \left(E_p(\mathbf{p}_{i,j}^c, \mathbf{W}_i^c) + E_d(\mathbf{r}_i^c, \mathbf{p}_{i,j}^c)\right) & \text{if } y_i^c = l^c \\ 0 & \text{if } y_i^c = l^0, \end{cases} \tag{2}$$

where E_r and E_p are the unary potentials for the root and part filters respectively. E_d encodes the deformation cost between the root $\mathbf{r}_i^c$ and each part $\mathbf{p}_{i,j}^c$. E_r, E_p and E_d are directly defined as in the original DPM [14]. Note that, in principle, if the candidate was generated by another detector, we could still extract the DPM model parameters from $\mathbf{W}_i^c$ and make use of this unary potential.

Pairwise Potentials. The pairwise potential ψ_p is a data-dependent smoothing term that encourages similar hypotheses to share the same object label. As in [21], we restrict our pairwise potential to take the form of a weighted mixture of Gaussian kernels, which can be expressed as

$$\psi_p(y_i^c, y_j^c|\mathbf{X}_i^c, \mathbf{X}_j^c) = \mu(y_i^c, y_j^c) \sum_{m=1}^{M} w^m k^{(m)}(\mathbf{f}_i^c, \mathbf{f}_j^c) \ , \tag{3}$$

where $\{w^m\}$ are the weights of the Gaussian kernels $\{k^{(m)}\}$, and μ is a label compatibility function. In particular, we make use of this function to encode a data-dependent Potts model, i.e., $\mu(y_i, y_j) = \mathbf{1}_{y_i \neq y_j}$.

The mixture of Gaussian kernels measures the appearance similarity between two object candidates. To this end, we use multiple feature types, as well as multiple kernel parameters. For each feature type $\mathbf{f}_s$, we construct a series of kernel functions of the form

$$k(\mathbf{f}_{i,s}^c, \mathbf{f}_{j,s}^c; t, \sigma_s) = \exp\left(-\frac{\|\mathbf{f}_{i,s}^c - \mathbf{f}_{j,s}^c\|^2}{2^t \sigma_s^2}\right) \ , \tag{4}$$

where σ_s is the minimum kernel width and t is an integer. We enumerate the value of t from 1 to T to define our series of kernels. Using kernels with different widths provides us with more flexibility in the representation of the similarity. The mixture of Gaussian kernels in Eq. 3 is then obtained by summing over all feature types and all values of t in each type. As will be discussed in Section 3.3, to avoid having to manually tune the weights $\{w^m\}$ of this mixture, we propose an efficient supervised learning procedure to estimate these weights.

Efficient co-detection. Given our fully-connected CRF model, we jointly detect the object instances in the input images by performing maximum posterior marginal inference. Following [21], we adopt a fast mean field approximation algorithm to compute the marginals. Given the current mean field estimates $\{Q_i\}$ of the marginals, the update equation can be written as

$$Q_i(y_i^c = l) \propto \exp\left(-\phi_u(y_i^c) - \sum_{l' \neq l} \sum_{j \neq i} Q_j(y_j^c = l')\psi_p(y_i^c, y_j^c)\right) . \tag{5}$$

Due to the mixture of Gaussian kernels form of the pairwise term, the updates can be computed in parallel by convolution with Gaussian kernels. This can be achieved efficiently by exploiting fast Gaussian filtering techniques, such as the permutohedral lattice-based method of [2].

After convergence, we obtain an (approximate) posterior distribution of object labels for each node (i.e., object candidate). To obtain the final co-detection results, we can then compute the most likely label for each object candidate, $\hat{y}_i^c = \arg\max_{y_i^c} Q_i(y_i^c)$. Furthermore, we can also exploit the mean field approximate marginal probability $Q_i(\hat{y}_i^c)$ as a detection score.

Note that, with our pairwise potential and since we treat each class separately, our CRF models a binary problem with a submodular energy function. As such, it could in principle be solved exactly by the graph-cut algorithm [5,20]. However, to achieve efficiency, the conventional graph-cut algorithm [5] relies on the sparse connectivity of the graph. As will be shown in Section 4, a graph-cut solution to our inference problem becomes significantly slower than our efficient filtering-based mean field solution when dealing with large densely connected random fields. Furthermore, note also that the MAP estimate from graph-cut does not provide a confidence score for the detection. Finally, despite that in this work we focus on a binary labeling problem, our formulation easily extends to the multi-class scenario.

3.3 Learning Object Similarity

Recall that our pairwise potentials encode the appearance similarity between two object candidates as a mixture of Gaussian kernels. To suitably adjust the weights of the mixture to the problem at hand, we can exploit training data and learn the weights that minimize the deviation from an ideal similarity measure. Here we formulate kernel weights estimation as a least-squares regression

problem, where the ground-truth (binary) similarity is directly defined by the compatibility of the labels of two object instances.

More specifically, we build a training set of object pairs, $\mathcal{D} = \{(\mathbf{X}_{i,1}, \mathbf{X}_{i,2}, s_i)\}$, where s_i is the ground-truth similarity, taking value 1 if $\mathbf{X}_{i,1}$ and $\mathbf{X}_{i,2}$ belong to the same class (excluding background) and 0 otherwise. The weights of the kernels can then be estimated by solving the optimization problem

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \left(s_i - \sum_{m=1}^{M} w^m k^{(m)}(\mathbf{f}_{i,1}, \mathbf{f}_{i,2}) \right)^2 , \tag{6}$$

where $\mathbf{w} = \{w^1, \cdots, w^M\}$ contains the weights of all kernels for all feature types. Note that this is a least-squares problem, and that its solution can therefore be obtained in closed-form.

To compute ground-truth similarities for category-level co-detection, we employ the following procedure. For each object class c, we first apply the same pre-trained object detector with high recall on the training images, and compute the intersection-over-union (IOU) of each detected bounding box with respect to the ground-truth bounding boxes. A detected bounding box is said to belong to class c if its maximum IOU w.r.t. ground-truth bounding boxes is larger than 50%. Otherwise, it is labeled as background. The training set $\mathcal{D}$ can then be constructed by collecting all possible pairs of detected bounding boxes and setting their similarity to 1 if they were both found to belong to class c, and 0 otherwise. In practice, the number of such pairs grows quickly, and we therefore randomly subsample the dissimilar pairs to build a balanced training set. Note that, as will be shown in our experiments, this procedure can also make use of instance-level labels when available, even if the final task remains category-level co-detection. In this scenario, two bounding boxes are considered similar only if they depict the same instance from the general category c.

4 Experiments

In this section, we study the effectiveness of our approach and compare it against state-of-the-art co-detection baselines.

4.1 Datasets and Setup

We evaluate our framework on several standard object co-detection datasets. These datasets include the Ford Car dataset of [3] and the Pedestrian dataset of [11], which provide category-level labels for the bounding boxes. Furthermore, we also employ the Human Co-detection (HCD) dataset of [24], which provides instance-level annotations of the bounding boxes. Note, however, that the task for HCD remains that of category-level co-detection, but, as suggested in [24], the instance-level annotations can be employed to better model object similarities.

In our experiments, we used the version 4 of DPMs [13], since it was also employed in [18,24]. This version provides one root and eight parts for each

object. For each object candidate, we computed a 59 dimensional Local Binary Patterns (LBP) feature and a 32 dimensional color histogram feature on the H channel of the HSV color-space. Note that the dimensionality of these features can be reduced using PCA to speed up inference. Our method has three hyper-parameters: the number of kernels, the widths σ_s and the pairwise weight α. These parameters were obtained by two-fold cross validation.

In our results, each bounding box is labeled based on the mean field approximate marginal probability of the object class. This lets us compute precision-recall curves, as opposed to a single point on these curves if we used the MAP estimate. We report average precision (AP) at category level following the evaluation metric in the PASCAL VOC challenge. For a bounding box to be considered correct, it must have at least 50% overlap with one of the ground truth bounding boxes in that image. This also has the advantage of making our results directly comparable to previously-reported ones. Therefore, for the baselines, we directly quote results reported in [3,18,24]. Our results for all experiments were averaged over 10 random training/test partitions.

We compare our method with the DPM baseline [14] and the following state-of-the-art co-detection approaches: 1) Object Co-detection [3]; 2) Multi-feature Joint Low-Rank Reconstruction [18]; 3) Human Co-detection and Labeling [24]. To study the effect of using different image features in our similarity kernels, we also consider two simpler versions of our approach, each of which uses only one feature type (either LBP [23] or color histograms). We refer to these two systems as LBP-CRF and Color-CRF, respectively, and to our full system as Joint-CRF.

4.2 Results and Discussion

Pedestrian Dataset. The Pedestrian dataset consists of 476 training images and 374 test images from two video sequences of street scenes acquired with a stereo setup. Each image has a resolution of 640×480 and contains multiple people. To evaluate our co-detection framework, we follow the same scenario as other co-detection work, which address the problem of jointly detecting people in a pair of images. Note that this dataset provides ground truth labels only for the left images in the stereo pairs. To mimic the stereo scenario, we therefore follow the same strategy as [18] and generate pseudo-stereo pairs by randomly drawing pairs of images that are no more than 3 frames apart in the left sequences. We generate 476 training pairs from the left training sequence and 300 test pairs from the left test sequence in this manner.

In Table 1, we report the results of our approach and the baselines on this dataset. Note that our approach significantly improves the results of the DPMs. More importantly, we also outperform all the baselines, even [24] that is specifically dedicated to the human co-detection case. This is also true for the single-feature versions of our model (LBP-CRF and Color-CRF). Combining these two features in our Joint-CRF model nonetheless lets us further improve performance. Sample co-detection results are given in Fig. 3. We also evaluate our method with random pairs as in [3] and achieve a similar performance, as shown in Table 1.

Table 1. Pedestrian co-detection: Comparison of our approach with state-of-the-art co-detection methods on the Pedestrian dataset

	Stereo Pairs		Random Pairs	
Methods	Ped(all)	Ped(h>120)	Ped(all)	Ped(h>120)
DPM [13]	59.7	55.4	59.7	55.4
Obj. Co-detection [3]	62.7	63.4	58.1	58.1
Robust Obj. Co-detection [18]	67.8	70.1	67.7	70.3
Human CoDeL [24]	74.4	73.8	-	-
LBP-CRF	77.04	79.43	-	-
Color-CRF	77.99	79.70	-	-
Joint-CRF	**78.73**	**81.25**	**77.7**	**80.43**

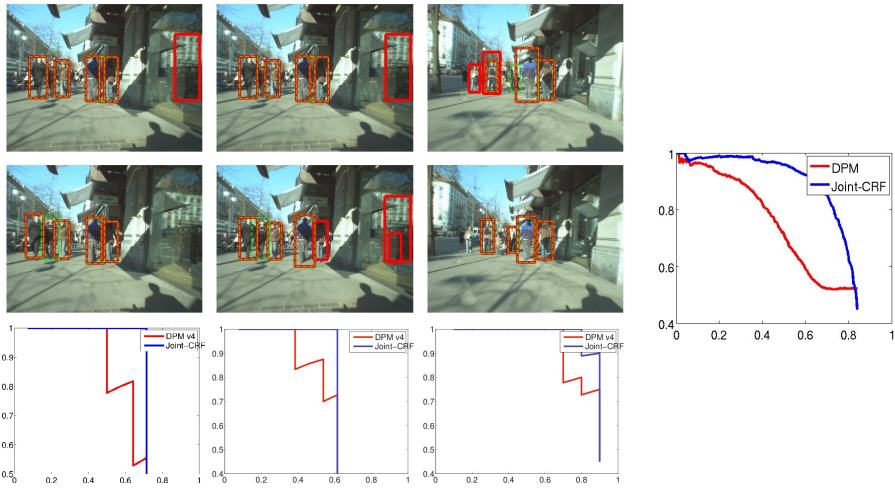

Fig. 3. Sample Results: Examples of our co-detection results on test pairs of the Pedestrian dataset. Top two rows: Input image pairs (Green dash: our results, Red solid: DPM results); Bottom row: Precision-recall curves of our method and DPM for the three image pairs. The precision-recall curves over all pairs are shown on the right.

We then study the quality of the similarity function learned from the training pairs using the method described in Section 3.3. To this end, in Fig. 4, we compare the similarity matrix obtained by applying the learned function to the candidates in one test pair with the corresponding ground-truth similarity computed from the correct labels (pedestrian vs background). Note that the predicted similarity depicts a similar pattern to the ground-truth one. This is further evidenced by the histogram that shows that pedestrian candidates have a high similarity score.

Ford Car Dataset. The Ford Car dataset consists of five scenes, each of which contains 86 stereo images. Each image has a resolution of 781×601 and depicts

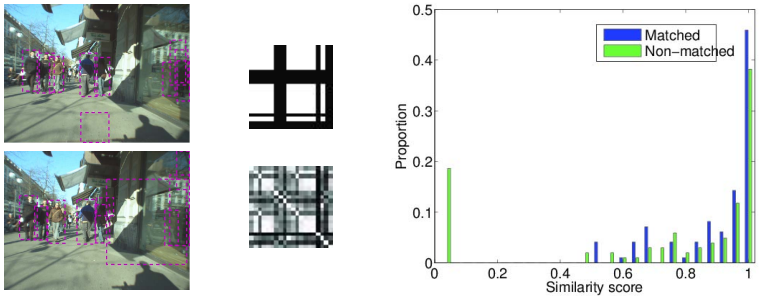

Fig. 4. Predicting similarity: Sample similarity matrix obtained by applying our learned similarity function to one test pair of the Pedestrian dataset. Left: Input image pair; Middle: (Top) target (ground truth) similarity matrix, (Bottom) learned similarity matrix (brighter means more similar); Right: Normalized histograms of similarity scores for matched and non-matched candidate pairs.

Table 2. Car co-detection: Comparison of our approach with state-of-the-art co-detection methods on the Ford Car dataset

	Stereo Pairs		Random Pairs	
Method	Ford(all)	Ford(h>80)	Ford(all)	Ford(h>80)
DPM [13]	49.8	47.1	49.8	47.1
Obj. Co-detection [3]	53.5	55.5	50.0	49.1
Robust Obj. Co-detection [18]	55	57.5	55.1	57.5
LBP-CRF	60.13	61.67	-	-
Color-CRF	59.44	60.45	-	-
Joint-CRF	**60.77**	**61.45**	**62.49**	**59.13**

multiple instances of cars at different scales and orientations. We made use of the 300 pseudo-stereo pairs provided by [3], which were generated in the same manner as described above for the Pedestrian dataset. Since no training pairs are provided, we extracted them in the same fashion, while ensuring no overlap with the test pairs. This resulted in a total of 410 training pairs.

The results of our approach and the baselines on this dataset are reported in Table 2. Note that, since it is dedicated to human co-detection, the method of [24] does not apply to this dataset. As in the Pedestrian case, our approach yields a significant performance improvement over the baselines. This is the case both with the single-feature models and with our Joint-CRF model. Sample co-detections are provided in Fig. 5. In addition, we also use random pairs, as in [3], for evaluation and obtain similar results. This shows that our method does not rely on the temporal information in the dataset.

Similarly to the Pedestrian case, in Fig. 6, we illustrate the quality of the learned similarity function by depicting the similarity matrix obtained when applying this function to one test pair. We can again see that the predicted similarity correctly reflects the ground-truth one.

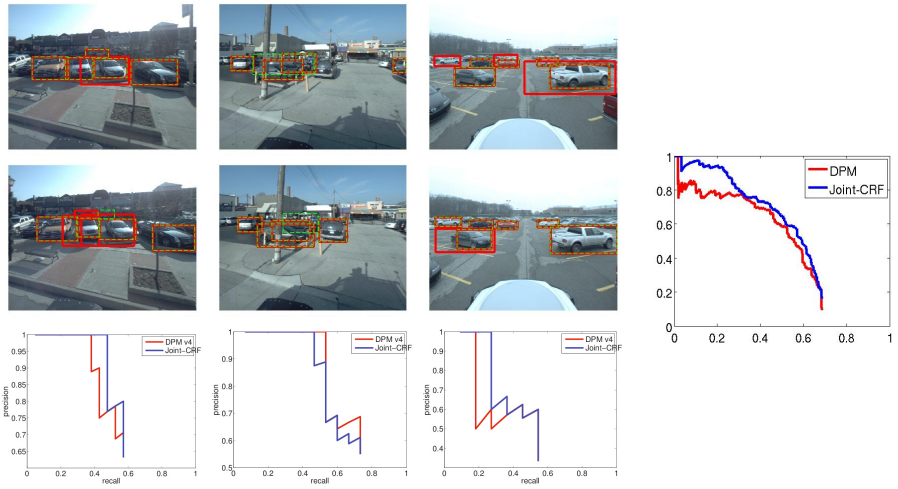

Fig. 5. Sample Results: Examples of our co-detection results on test pairs of the Ford Car dataset. Top two rows: Input image pairs (Green dash: our results, Red solid: DPM results); Bottom row: Precision-recall curves of our method and DPM for the three image pairs. The precision-recall curves over all pairs are shown on the right.

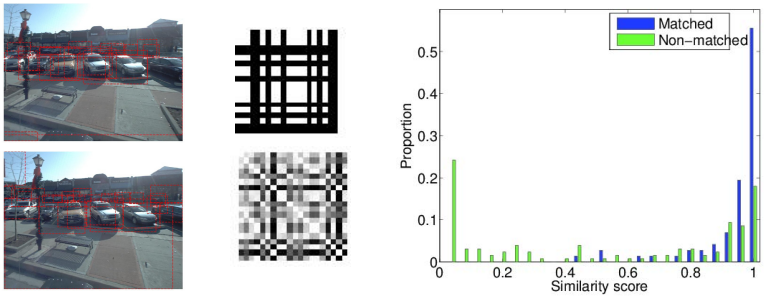

Fig. 6. Predicting similarity: Sample similarity matrix obtained by applying our learned similarity function to one test pair of the Ford Car dataset. Left: Input image pair; Middle: (Top) target (ground truth) similarity matrix, (Bottom) learned similarity matrix (brighter means more similar); Right: Normalized histograms of similarity scores for matched and non-matched candidate pairs.

Human Co-Detection Dataset. The HCD [24] comprises 387 images separated into 26 sets. Each image may contain multiple people, and the appearance of these people is consistent within one set. As opposed to the Ford Car and Pedestrian datasets where only two images with relatively small viewpoint difference are employed for co-detection, all the images in one set of HCD are considered simultaneously and typically depict large viewpoint changes. For our experiments, we followed a leave-five-sets-out strategy, which amounts to using roughly 80% of the images as training data and the remaining 20% (coming from

Table 3. Human co-detection: Comparison of our approach with state-of-the-art co-detection methods on the HCD dataset

	HCD Dataset
DPM	69.64
Human CoDeL [24]	74.94
LBP-CRF	78.81
Color-CRF	79.19
Joint-CRF	**79.41**

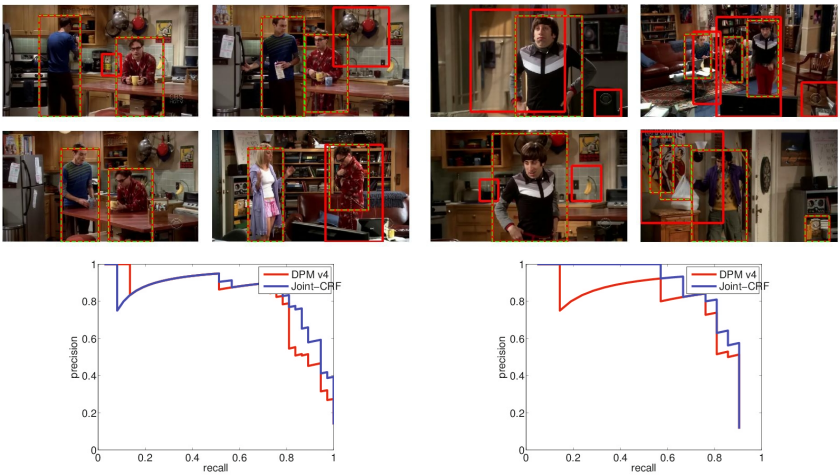

Fig. 7. Sample Results: Examples of our co-detection results on two sets from the Human Co-detection dataset. Top: Input image set overlaid with detection output (Green dash: our results, Red solid: DPM results); Bottom: Precision-recall curves of our method and DPM for the two sets.

five independent sets) as test images. For this dataset, we employed the provided instance-level labels to learn our similarity function. Note however that the main task remains category-level (human) co-detection.

We report our results on this dataset in Table 3. Note that the only reported results on HCD are those of [24]. As for the other datasets, we outperform the baselines, whether using a single feature type or multiple ones. Fig. 7 depicts some of our co-detection results on HCD. For this dataset, we also compare our results with those of a sparse CRF constructed by connecting only the first k nearest neighbors (based on our similarity) of each node, with k ranging from 1 to 50. The best F1 score of this sparse CRF (over all values of k) is 80.45%. This is clearly outperformed by our F1 score of 85.3%.

In Fig. 8, we also show the predicted similarity function for one of the sets in the dataset. Note that, because of the instance-level annotations, the similarity matrix is more complex than before. Nonetheless, our predicted similarity matrix still yields a good approximation of the ground-truth one.

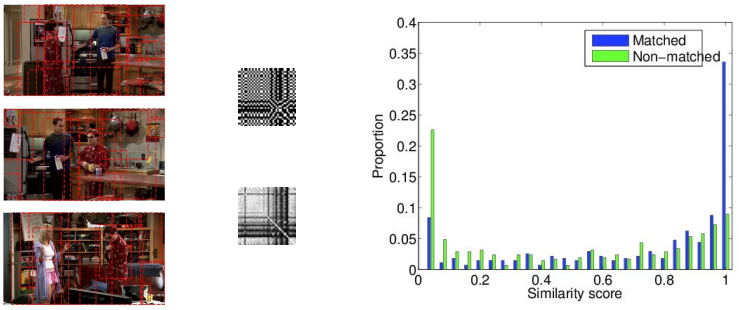

Fig. 8. Predicting similarity: Sample similarity matrix obtained by applying our learned similarity function to one test set in the Human Co-Detection dataset. Left: Input images (three examples); Middle: (Top) target (ground truth) similarity matrix, (Bottom) learned similarity matrix (brighter means more similar); Right: Normalized histograms of similarity scores for matched and non-matched candidate pairs.

Scaling up. As mentioned earlier, inference in our model could in principle also be performed using the Graph-cut algorithm [5]. Here, we study the scalability of both inference strategies with respect to the number of images considered jointly at test time. To this end, we build two fully-connected CRF models with different numbers of test images from the HCD dataset. The first CRF has 330 nodes. In this case, our mean field filtering inference takes 6.5 seconds and the Graph-cut only takes 0.6 second (this includes the time to compute the potential functions). However, when the size of the CRF increases by a factor 10, our method takes 8.5 seconds, which is only mildly slower compared to the previous setting. In contrast, the Graph-cut spends 30 seconds in potential calculation and 60 seconds in inference. This shows that the Graph-cut algorithm does not scale up to larger test sets for fully connected graphs, and thus confirms our choice of inference strategy.

5 Conclusion

In this paper, we have introduced a formulation of object co-detection from a pool of images that expresses the problem as inference in a fully-connected CRF whose nodes represent object candidates. We have then shown that modeling the similarity between pairs of candidates as a weighted mixture of Gaussian kernels allowed us to efficiently perform inference in our graph, while yielding an effective representation for co-detection. Our experimental evaluation has demonstrated that our approach could effectively leverage the information in multiple images to improve detection accuracy, thus outperforming existing co-detection techniques on benchmark datasets. In the future, we intend to study how our framework can be applied to jointly co-detecting different object categories, thus leveraging the collective power not only of multiple images, but also of multiple classes.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR (2010)
2. Baek, J., Adams, A., Dolson, J.: Lattice-based high-dimensional gaussian filtering and the permutohedral lattice. JMIV (2013)
3. Bao, S.Y., Xiang, Y., Savarese, S.: Object co-detection. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 86–101. Springer, Heidelberg (2012)
4. Barinova, O., Lempitsky, V., Kohli, P.: On detection of multiple object instances using hough transforms. PAMI (2012)
5. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. PAMI 23(11) (2001)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
8. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV (2009)
9. Dickinson, S.J., Leonardis, A., Schiele, B., Tarr, M.J.: Object Categorization: Computer and Human Vision Perspectives. Cambridge University Press (2009)
10. Endres, I., Hoiem, D.: Category independent object proposals. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 575–588. Springer, Heidelberg (2010)
11. Ess, A., Leibe, B., Gool, L.V.: Depth and appearance for mobile scene analysis. In: International Conference on Computer Vision, ICCV 2007 (October 2007)
12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)
13. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Discriminatively trained deformable part models, release 4, `http://people.cs.uchicago.edu/~pff/latent-release4/`
14. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE PAMI (2010)
15. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. CVIU 114, 712–722 (2010)
16. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. ArXiv 1311.2524 (2013)
17. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. JMLR 12 (2011)
18. Guo, X., Liu, D., Jou, B., Zhu, M., Cai, A., Chang, S.F.: Robust Object Co-detection. In: CVPR (2013)
19. Hoiem, D., Efros, A.A., Hebert, M.: Putting Objects in Perspective. IJCV 80, 3–15 (2008)
20. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Trans PAMI 26(2) (2004)
21. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS (2011)
22. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV (1999)
23. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. PAMI 24(7) (2002)

24. Shi, J., Liao, R., Jia, J.: CoDeL: An efficient human co-detection and labeling framework. In: ICCV (2013)
25. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. IJCV (2013)
26. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV (2009)
27. Vineet, V., Warrell, J., Torr, P.H.S.: Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 31–44. Springer, Heidelberg (2012)
28. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57(2) (2004)
29. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: NIPS (2006)
30. Zhang, Y., Chen, T.: Efficient inference for fully-connected crfs with stationarity. In: CVPR (2012)