Robust Object Co-Detection

Xin Guo[†], Dong Liu[‡], Brendan Jou[‡], Mojun Zhu[‡], Anni Cai[†], Shih-Fu Chang[‡]
[†]Beijing University of Posts and Telecommunications, Beijing, China
[‡]Dept. of Electrical Engineering, Columbia University, New York, NY, USA
{guoxin, annicai}@bupt.edu.cn, {dongliu, bjou, sfchang}@ee.columbia.edu, mz2330@columbia.edu

Abstract

Object co-detection aims at simultaneous detection of objects of the same category from a pool of related images by exploiting consistent visual patterns present in candidate objects in the images. The related image set may contain a mixture of annotated objects and candidate objects generated by automatic detectors. Co-detection differs from the conventional object detection paradigm in which detection over each test image is determined one-by-one independently without taking advantage of common patterns in the data pool. In this paper, we propose a novel, robust approach to dramatically enhance co-detection by extracting a shared low-rank representation of the object instances in multiple feature spaces. The idea is analogous to that of the well-known Robust PCA [28], but has not been explored in object co-detection so far. The representation is based on a linear reconstruction over the entire data set and the low-rank approach enables effective removal of noisy and outlier samples. The extracted low-rank representation can be used to detect the target objects by spectral clustering. Extensive experiments over diverse benchmark datasets demonstrate consistent and significant performance gains of the proposed method over the state-of-the-art object co-detection method and the generic object detection methods without co-detection formulations.

1. Introduction

Given an image and a target object category, the goal of object detection is to localize the instance of the given category within the image, often up to bounding box precision. The task is often challenging because the visual appearance of objects is often diverse due to occlusions, cluttered background, illumination and viewpoint changes.

The classical approach to object detection is to train object detectors from manually labeled bounding boxes in a set of training images and then apply the detectors on the individual test images. Despite previous success, this strategy only focuses on obtaining the best detection result within *one image* at a time and fails to leverage the consistent object appearance often existent when there are multiple related images. A promising alternative, called object co-detection [2], is to simultaneously identify "similar" objects in a set of related images and use intra-set appearance consistency to mitigate the visual ambiguity.

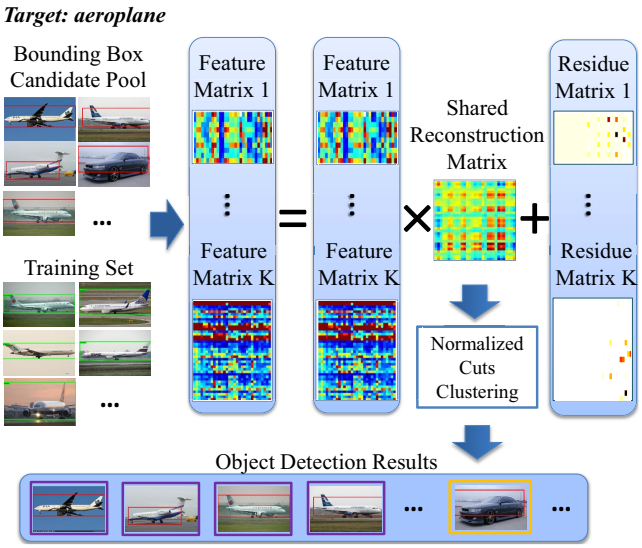

Figure 1. Illustration of robust object co-detection. Given a pool of automatically detected candidate regions and the training bounding box set, we represent them using K different features. For each feature matrix, we perform linear reconstruction, representing each bounding box as a linear combination of other bounding boxes where the resulting coefficient matrix measures the mutual dependency of bounding boxes. We derive a shared low-rank reconstruction matrix from the K reconstructions while removing the noisy and outlying bounding boxes in each feature matrix in a sparse residue matrix. The low-rank reconstruction coefficient matrix is then fed into Normalized Cuts clustering to yield co-detection results.

The previous method for object co-detection measures the consistency of the object appearances across images using pairwise matching, *i.e.*, objects belonging to the same category are expected to have high visual similarity. However, there are two main issues with this approach. First, the requirement for "pairs" means that it only seeks to model the two object instances at a time, not accounting for the

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.412

structure of the entire object space that may exist when there are more than two samples. This results in a direct loss of structure information when extending beyond pairwise relations, defeating the very purpose and advantage of co-detection. Second, previous co-detection methods have not considered noise and outlier object instances often caused by significant content variation.

We propose a novel object co-detection method which addresses these two issues and show empirically that improved detection results follow naturally from exploiting information from multiple images during detection. The overall procedure for our proposed co-detection method is illustrated in Figure 1. Given a target object category and a training corpus with bounding box annotations, we first train several state-of-the-art object detectors so that diverse appearances of the target object can be covered and the family of detectors can collectively reach a high recall in detection accuracy. These detectors are then applied to the test images to obtain an initial bounding box candidate pool. With the bounding boxes from the training set and the initial candidate pool over the test images, we extract K low-level features from each of them. For each feature, we perform a linear reconstruction task to represent each bounding box as a linear combination of other bounding boxes such that the reconstruction coefficients represent the dependency of one bounding box to the others. We seek to find a shared low-rank reconstruction coefficient matrix across these K reconstructions that captures the global structure of the object space while removing noise and outliers in each feature space via a sparse residue matrix. We formulate the problem as a constrained nuclear and $\ell_{2,1}$ norm minimization problem and use the Augmented Lagrange Multiplier (ALM) [14] method for efficient optimization.

Notably, our method leverages unlabeled data and is similar to semi-supervised learning. However, the difference is that the unlabeled data is not given arbitrarily, but corresponds to potential bounding boxes generated by multiple detectors. On such candidate data, the low-rank assumption is more likely to hold. The use of low-rank constraints on the coefficient matrix is particularly important for discovering the mutual dependence that may exist between bounding boxes, which we refer to as the "global structure". To capture this structure on bounding boxes, we assume that the reconstruction coefficient vectors is dependent on each other. In many cases, due to the intrinsic complexity of object appearance, it is often impossible to find a single feature to accurately measure the mutual dependency among objects. While different features may yield different low-rank coefficient matrices, a shared low-rank coefficient matrix is necessary because it captures object dependency across these features and in so doing, ensures robustness. Noise and outliers from each feature space can also be removed via a sparse residue matrix which reduces ambiguity that may have been introduced by each feature. Our experiments on benchmark datasets used by [2] as well as on PASCAL VOC 2007 and 2009 show consistent and significant margins of improvement over generic object detectors using little prior and the state-of-the-art object co-detector.

2. Related Work

Object detection is a longstanding challenge in the computer vision community. Perhaps most notably, sliding window classifiers have gained enormous popularity as they are especially well suited for faces [27], and rigid objects like pedestrians [9] and cars [22]. In this setting, typically, a binary classifier is first trained on annotated object instances and then evaluated using a uniform sampling of possible locations and scales in each test image followed by post-processing step to find objects, such as non-maximum suppression. Felzenszwalb *et al.* [12] proposed a deformable part-based model which assumes that an object is composed by several deformable components whose positions are treated as latent variables and are applied toward a latent support vector machine (SVM) for detection. Recently, Malisiewicz *et al.* [18] proposed the use of an ensemble of Exemplar-SVMs for object detection, where a SVM is trained on a single positive instance and many negatives during training. This better captured the specific visual appearance of each positive instance and performed competitively against many of the part-based detectors with growth out of [12]. However, these works are largely optimized on taking only one image into consideration at a time and neglect the collective information when there is a corpus of images available at test time.

The most relevant work to our proposed method is the co-detection work of Bao *et al.* [2]. Similar to our setting, they seek to detect the target objects that simultaneously appear in a set of images. Specifically, they represent an object category using part-based object representations and measure appearance consistency between objects by pairwise similarity matching. In their approach, the information from multiple images are combined through an energy-based formulation that models both within-image and cross-image similarities. Again, this ignores the global structure in the object space and also assumes the absence of large variance noise. In contrast, we focus on collectively discovering global structure from an object bounding box pool and concurrently removing outliers, which we believe leads to robust object co-detection, able to handle noise. Our method is also closely related to image co-segmentation [23, 20], which performs simultaneous segmentation of the shared foreground regions in a set of images but does not attempt to recognize the object identity. Contrarily, co-detection aims at assigning a category label to each detection object.

Our work is motivated by Robust PCA [28] and recent

low rank matrix recovery works [6, 5]. In particular, Liu *et al.* [15] proposed the low-rank representation method which can be used to discover the underlying subspace structures by imposing the low-rank constraint on the representation coefficient matrix while using $\ell_{2,1}$-norm to remove outliers. We have recently seen several successful similar formulations of this toward applications including image segmentation [8] and saliency detection [24]. Our method is distinct in that we develop a low-rank coefficient matrix that is shared over multiple reconstructions derived from different features. We note that related work can also be found in multi-task joint sparse representation [30], but it seeks to find stable training images across multiple features to classify test images rather than using them to discover the global structure as we do toward object localization in images.

3. Robust Object Co-Detection

In this section, we introduce our robust object co-detection method based on multi-feature low-rank reconstruction. We first present how we generate an initial pool of candidate bounding boxes of the target object and then describe our problem formulation. Finally, we explain how to use a learned low-rank coefficient matrix for co-detection.

3.1. Bounding Box Candidate Pool Generation

Exhaustive window scanning will generate a massive number of bounding boxes that dramatically increases the computational burden of the object detector. Therefore, an initial bounding box generation procedure is necessary to prune the windows that do not contain any target object. Given a target object category and its associated training bounding boxes, we train two kinds of object detectors: Deformable Part-based Model (DPM) [12] and Ensemble of Exemplar-SVMs (ESVMs) [18]. These two detectors complementarily model the object appearance via their native choices of feature representation. A similar bounding box candidate pool generation method was adopted in [2] using DPM. We apply the detectors on each test image and select the top B bounding boxes with the highest detection scores as the potential localizations on that image. We set B to be twice the average number of bounding boxes in the training images[1]. Because we have two detectors, there are $2B$ bounding box suggestions for each test image. After removing the duplicate bounding boxes with non-maximum suppression, we obtain an initial bounding box pool with a high recall. We note that other bounding box pool generation methods, such as objectness detection [1] may be also considered as alternatives to these two detectors.

3.2. Problem Formulation

Given an object category, suppose we have l training bounding boxes and u potential bounding boxes from the initial bounding box pool. We extract low-level features from each bounding box and obtain a feature matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_{l+u}]$, where $\mathbf{x}_i \in \mathbb{R}^m$ is the feature vector of the i-th bounding box ($i = 1, \ldots, l+u$) with m denoting the dimensionality of the feature vector. We begin by considering the following linear reconstruction problem[2]:

$$X = XZ + E, \tag{1}$$

where $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_{l+u}] \in \mathbb{R}^{(l+u)\times(l+u)}$ is the reconstruction coefficient matrix with $\mathbf{z}_i \in \mathbb{R}^{l+u}$ denoting the reconstruction coefficient vector of bounding box $\mathbf{x}_i$. Notably, the j-th entry in vector $\mathbf{z}_i$ is the contribution of the bounding box $\mathbf{x}_j$ in reconstructing the bounding box $\mathbf{x}_i$, and measures the mutual dependence between $\mathbf{x}_i$ and $\mathbf{x}_j$. E is the reconstruction residue matrix of the given feature matrix X.

There are two issues associated with the above linear reconstruction. First, it finds the reconstruction coefficient vector for each bounding box individually, and hence does not take into account the global structure of the bounding boxes. Second, it cannot remove undesired noise and outliers which may degrade detection performance. To solve the above two issues, we consider the following matrix decomposition problem:

$$\begin{aligned} \min_{Z,E} \quad & \operatorname{rank}(Z) + \lambda \|E\|_{2,1}, \\ \text{s.t.} \quad & X = XZ + E, \end{aligned} \tag{2}$$

where $\operatorname{rank}(Z)$ is the rank of the matrix Z, $\|E\|_{2,1} = \sum_{i=1}^{l+u} \sqrt{\sum_{j=1}^{m} (E_{i,j})^2}$ is the $\ell_{2,1}$-norm, and $\lambda \geq 0$ is a tradeoff parameter balancing the two competing terms.

The minimization of $\operatorname{rank}(Z)$ forces the reconstruction coefficient matrix to have the lowest rank possible. As a result, the reconstruction coefficient vectors of different bounding boxes influence each other in such a way as to encourage bounding boxes to be linearly spanned by only a few bases. The matrix Z then represents the global structure of the bounding boxes. The second term $\|E\|_{2,1}$ ensures that a small number of columns are non-zero, restricting the amount of noise that leaks into the feature matrix. By removing E from X, the feature representations of the bounding boxes become more compact, reducing potential ambiguity in the detection process.

The objective function of (2) is based on a single feature modality. In general, we require more than one feature to discover the global structure of the objects given their diverse visual appearance. A more promising alternative is to find a reconstruction coefficient matrix shared across multiple features, whose entries can more precisely reflect

[1]Note that studying the optimal choice of B is a legitimate research problem but not the main focus of this paper.

[2]Linear reconstruction has been successfully applied in several recent works on sparse representation [29], subspace clustering [15], etc. Indeed, our method can be extended to the non-linear case, *e.g.*, graph-based reconstruction.

the degree of contribution from features on the mutual dependence between any two bounding boxes. Given K total features, for the k-th feature where $k = 1, \ldots, K$, let $X^k = [\mathbf{x}_1^k, \ldots, \mathbf{x}_{l+u}^k]$ be the feature matrix of all the bounding boxes where $\mathbf{x}_i^k \in \mathbb{R}^{m_k}$ is the feature vector of the i-th bounding box ($i = 1, \ldots, l+u$) with m_k being the dimensionality of the k-th modality. We consider the objective:

$$\begin{aligned} \min_{Z,E^k} \quad & \operatorname{rank}(Z) + \lambda \sum_{k=1}^{K} \|E^k\|_{2,1}, \\ \text{s.t.} \quad & X^k = X^k Z + E^k, \;\; k = 1, ..., K, \end{aligned} \tag{3}$$

where E^k is the residue matrix removed from X^k. Note that the coefficient matrix Z is shared across K features.

The above optimization problem is difficult to solve due to the discrete nature of the rank function. Instead, we focus on the following tractable convex objective which is a good surrogate of the original optimization problem:

$$\begin{aligned} \min_{Z,E^k} \quad & \|Z\|_* + \lambda \sum_{k=1}^{K} \|E^k\|_{2,1}, \\ \text{s.t.} \quad & X^k = X^k Z + E^k, \;\; k = 1, \ldots, K, \end{aligned} \tag{4}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, *i.e.*, the sum of the singular values.

3.3. Object Co-Detection with Matrix Z^*

After solving for the global structure matrix Z^* from (4), we can use it to simultaneously detect all target objects from a bounding box collection consisting of the training annotations and the initial bounding box pool from Section 3.1. We accomplish this task via a clustering procedure which partitions the bounding boxes so that each cluster contains objects with the same visual appearance. Since the coefficient matrix Z^* inherently captures the mutual dependence of the bounding boxes, it is natural to employ it as an affinity measure for clustering. To ensure the symmetric property of affinity matrices, we convert Z^* into a symmetric affinity matrix W via the relation [15]:

$$W = \frac{1}{2}\big(|(Z^*)^\top| + |Z^*|\big). \tag{5}$$

Using this affinity matrix, we employ Normalized Cuts [25] to segment bounding boxes into N clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$. We then define the detection score s_i for the i-th test bounding box ($i = l+1, \ldots, l+u$) as:

$$s_i = \frac{\max\{|\mathcal{P}(\mathcal{C}_{I_i})|, 1\}}{\max_{1 \le q \le N} |\mathcal{P}(\mathcal{C}_q)|} \times \frac{\sum_{j \in \mathcal{P}(\mathcal{C}_{I_i})} W_{ij}}{\max\{|\mathcal{P}(\mathcal{C}_{I_i})|, 1\}}, \tag{6}$$

where I_i is an indicator specifying the index of the cluster which the i-th test bounding box belongs to, $\mathcal{P}(\mathcal{C}_q)$ is the set of positive training bounding boxes in cluster $\mathcal{C}_q$, and $|\cdot|$ denotes the cardinality of a set. Here we use $\max\{|\mathcal{P}(\mathcal{C}_{I_i})|, 1\}$ operator to ensure that clusters that have no positive samples do not cause division by zero. Though the numerator of left-hand-side and denominator of right-hand-side term will cancel out, we present Eq.(6) so as to clearly show how scores arise. The first right-hand-side term is a weighting term that gives clusters with greater number of positive training samples higher weight. This is accomplished by dividing the number of positive training samples in the same cluster as the i-th sample by the highest number of per-cluster positive training samples across all clusters. The result is that clusters with more positive training samples have higher voting power and thus, the scores for test samples in those clusters are likely to have higher weight. The second term is simply the average of affinities between the i-th sample and all positive training instances in the same cluster. With these scores on test bounding boxes, we can then obtain a rank list in which the highest positive detections are ranked in the top positions. These top ranking bounding boxes correspond to the result of our co-detection for that respective object category.

4. Optimization Procedure

The problem is a mixed nuclear norm and $\ell_{2,1}$-norm optimization problem, which can be easily solved by the Augmented Lagrange Multiplier (ALM) [14] method.

First, we convert the problem in (4) into the following equivalent form:

$$\begin{aligned} \min_{Z,E^k,J} \quad & \|J\|_* + \lambda \sum\nolimits_{k=1}^{K} \|E^k\|_{2,1}, \\ \text{s.t.} \quad & X^k = X^k Z + E^k, \;\; k = 1, ..., K, \\ & Z = J. \end{aligned} \tag{7}$$

This problem can be solved by the Augmented Lagrange Multiplier (ALM) method [14], which minimizes the following augmented Lagrange function:

$$\begin{aligned} & \min_{Z,E^k,J,Y^k,U} \|J\|_* + \lambda \sum\nolimits_{k=1}^{K} \|E^k\|_{2,1} \\ & + \sum\nolimits_{k=1}^{K} \big\langle Y^k, X^k - X^k Z - E^k \big\rangle + \big\langle U, Z - J \big\rangle \\ & + \frac{\mu}{2}\big(\sum\nolimits_{k=1}^{K} \|X^k - X^k Z - E^k\|_F^2 + \|Z - J\|_F^2\big), \end{aligned} \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the trace of inner product, $Y^k (k = 1, \ldots, K)$ and U are the Lagrange multipliers, and $\mu > 0$ is a penalty parameter. For fast convergence speed, we use inexact ALM to solve (8) and resulting optimization procedure is found in Algorithm 1. Step 4 is solved by adopting singular value thresholding operator [4] and Step 6 is solved via the analytic solution in [16].

We implemented Algorithm 1 in MATLAB on a machine equipped with a 12-Core Intel Xeon E5649 processor, 2.65 GHz CPU and 48 GB memory, and have generally observed that the iterative optimization converges fast. For example, in detecting the "aeroplane" category on PASCAL VOC 2007 dataset (see Section 5.2), a single iteration of the

Algorithm 1 Solving (8) by Inexact ALM

1: **Input:** multi-feature matrix X^k $(k = 1, \ldots, K)$, parameter λ
2: **Initialize:** $Z = J = 0, E^k = 0, \{Y^1, \ldots, Y^K, U\} = 0, \mu = 10^{-6}, \mu_{\max} = 10^{10}, \rho = 1.1$
3: **while** not converged **do**
4: Update J by
$J = \arg\min \frac{1}{\mu}\|J\|_* + \frac{1}{2}\|J - Z - U/\mu\|_F^2$
5: Update Z by
$Z = (I + \sum_{k=1}^{K} (X^k)^\top X^k)^{-1} (\sum_{k=1}^{K} ((X^k)^\top X^k - (X^k)^\top E^k) + J + (\sum_{k=1}^{K} (X^k)^\top Y^k - U)/\mu)$
6: Update E^k by
$E^k = \arg\min \frac{\lambda}{\mu}\|E^k\|_{2,1} + \frac{1}{2}\|E^k - (X^k - X^k Z + \frac{Y^k}{\mu})\|_F^2$
7: Update the multipliers
$Y^k = Y^k + \mu(X^k - X^k Z - E^k), \;\; k = 1, \ldots, K$
$U = U + \mu(Z - J)$
8: Update parameter μ by $\mu = \min(\rho\mu, \mu_{\max})$
9: Check convergence conditions: $\forall\, k = 1, \ldots, K$,
$X^k - X^k Z - E^k \to 0$
$Z - J \to 0$
10: **end while**
11: **Output:** $Z, X^k, E^k, \;\; k = 1, \ldots, K$

optimization of Steps 3 through 10 finishes within about 5 seconds. Since each optimization sub-problem in Algorithm 1 will monotonically decrease the objective function, the algorithm is guaranteed to converge.

5. Experiments

We evaluate our proposed method on several benchmark object detection datasets. We compared the following:

- Deformable Part-based Models (DPM) [12].
- Ensemble of Exemplar-SVMs (ESVMs) [18].
- Multi-Feature Matching (MFM): We first generate an initial bounding box pool through DPM and ESVMs, then rank all the candidate bounding boxes based on their average similarity with respect to the l training bounding boxes. The similarity is measured from multiple features[3].
- Multi-feature joint Sparse Reconstruction (MSR): The ℓ_1-norm is applied to the coefficient matrix shared over multiple features, *i.e.*, it does not capture the global structure across the objects.
- Single-feature Low-Rank Reconstruction (SLRR): We report the detection performance based on the each individual feature modality.
- Our method called Multi-feature joint Low-Rank Reconstruction (MLRR).
- Several other reported state-of-the-art object detection/co-detection methods [2, 10, 13, 31].

[3]For the j-th candidate bounding box, its similarity is calculated as $\bar{s}_j = \frac{1}{lK}\sum_{i=1}^{l}\sum_{k=1}^{K} s_{ij}^k$, where $s_{ij}^k = \exp(-d(\mathbf{x}_i^k, \mathbf{x}_j^k)/\sigma^k)$ is the Gaussian similarity based on the k-th feature modality, $d(\mathbf{x}_i^k, \mathbf{x}_j^k)$ is the χ^2 distance between $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$, and σ^k is the mean value of all pairwise χ^2 distances between the candidate and training bounding boxes.

We note that there are some recent works that using strong auxiliary information such as shape masks [7] and image contextual cues [26]. We do not include these methods into our comparison, but emphasize that our detection framework is applicable to any generic feature and outperforms the generic detection methods with using any of those priors or contexts. Moreover, to ensure fairness, we download and use the DPM models of PASCAL VOC 07 that yield the same results reported in [12, 18].

We extract three kinds of features from each bounding box including SIFT Bag-of-Words (BoW) [17], Gabor [19], and LBP [21] features. For SIFT BoW, we extract densely-sampled SIFT descriptors every 8 pixels with a patch size of 16×16. We then train a codebook with $1,024$ codewords and quantize the descriptors in each bounding box into a $1,024$-dimension histogram. For the Gabor feature, we partition each bounding box into 2×2 blocks and apply a set of Gabor filters over 4 scales and 6 orientations in each block. From the filter's response in each block, we use the mean and standard deviation as the feature descriptor. The resulting feature vector is 192 dimensions. For the LBP feature, we generate LBP codes using 8 neighbors on a circle of unit radius and obtain a 59-dimensional feature vector.

We note that our method has two parameters: the trade-off parameter λ and the number of clusters N. To determine the appropriate value for each parameter, we varied the value of λ on the range of $\{10^{-3}, 10^{-2}, \ldots, 10^{0}\}$, and cluster number N discretely between $\{10, 15, \ldots, 60\}$. We then choose the best parameter values based on validation performance. For the parameter setting of other competed methods, we follow the original suggestion parameter setting strategies and choose the best parameter values based on validation performance.

Following the evaluation method in PASCAL VOC challenge, a predicted bounding box is considered correct if it overlaps more than 50% with the ground-truth bounding box, otherwise it is considered a false detection. The average precision (AP) is computed from the precision/recall curve and is an approximation of the area under this curve. The mean AP (mAP) measures the mean of the APs over all categories of the dataset.

5.1. Comparison with [2]

Here, we distinguish our method from the recently proposed object co-detection method in [2]. We perform comparisons on two datasets used by the work: Ford Car [2] and Pedestrian dataset [11]. The Ford Car dataset consists of 430 images in five scenes and Pedestrian dataset contains 490 training frames and 354 test frames from two video sequences of the shopping street.

Since the co-detection method in [2] is based on pairwise matching, it selects various image pairs as the test set for performance evaluation. There are two kinds of image pairs they selected for testing: stereo pair and random pair.

Statistics	Stereo Pair	#unique images	141 ± 1	141 ± 1	285 ± 3	285 ± 3
		#bounding boxes per image	10.28 ± 0.01	8.4 ± 0.01	8.41 ± 0.04	5.16 ± 0.05
		average recall (%)	70.10 ± 1	74.43 ± 1	76.66 ± 1.3	85.12 ± 1.1
	Random Pair	#unique images	138 ± 2	138 ± 2	273 ± 2	273 ± 2
		#bounding boxes per image	10.24 ± 0.04	8.47 ± 0.05	8.47 ± 0.02	5.17 ± 0.01
		average recall (%)	70.06 ± 0.19	74.28 ± 0.05	76.83 ± 0.18	84.31 ± 0.2
Datasets & Methods			Ford Car (all)	Ford Car (h>80)	Pedestrian (all)	Pedestrian (h>120)
AP (%)	Stereo Pair	DPM [2]	49.8	47.1	59.7	55.4
		Co-Detector [2]	53.5	55.5	62.7	63.4
		Ours (MLRR)	**55.0 ± 0.1**	**57.5 ± 0.1**	**67.8 ± 0.9**	**70.1 ± 1.1**
	Random Pair	DPM [2]	49.8	47.1	59.7	55.4
		Co-Detector [2]	50.0	49.1	58.1	58.1
		Ours (MLRR)	**55.1 ± 1.4**	**57.5 ± 1.2**	**67.7 ± 1.3**	**70.3 ± 1.5**

Table 1. **Top**: Statistics of images pairs based on two different sampling strategies. **Bottom**: Performance comparison (AP %) on Ford Car and Pedestrian datasets.

The stereo image pairs are obtained from a stereo camera, meaning most images contain matching objects. The random image pairs are randomly selected from the dataset, where many the pairs contain few or no matching objects. Specifically, they select 300 and 200 image pairs for Ford Car and Pedestrian datasets under the two settings. The authors only provide 354 test stereo pairs for Ford Car dataset while the other pairs are not publicly available. To ensure a similar setting, we select 300 stereo pairs from the 354 available stereo pairs and select 300 random pairs from the whole dataset for testing on the Ford Car dataset.

For the Pedestrian dataset, since there are not any test pairs available, we follow the same stereo pair generation method as in the released pairs of the Ford Car dataset. We select 200 stereo pairs from test frames with the constraint that each pair consists of two frames whose frame interval is at most three within the video sequence. In addition, 200 random pairs are randomly selected from the test frames. Remaining images from Ford Car and the original training frames of Pedestrian dataset are used as training data. The experiments are repeated three times and the average performance and standard deviation are calculated. In each run, we perform two-fold cross validation on the training set to determine the best parameters for each method.

The comparison of results is shown in Table 1 and we see that our method outperforms the object co-detection method by a large margin on both datasets under different test pair sampling strategies. We see that the average margin of improvement for the co-detector of [2] from per-image detection is only about 5.77% and 0.82% for stereo and random pair, respectively, but we push it to 9.60% and 9.65%. Figure 2 shows example incorrect bounding boxes successfully removed by our method (corresponding to bounding boxes with non-zero columns in the residue matrix).

5.2. Experiments on PASCAL Datasets

We also evaluated our proposed method on two other benchmark object detection datasets: PASCAL VOC 2007 and VOC 2009.

PASCAL VOC 2007 dataset. This dataset contains $9,963$ images over 20 categories which are divided into "train", "val" and "test" subsets, *i.e.*, 25% for training ($2,501$ images), 25% for validation ($2,510$ images) and 50% for testing ($4,952$ images). For each method, we select the best parameter based on the validation performance on the validation set "val". The average recall rate across the 20 categories in the bounding box candidate pool is 59.8% and there are on average 6.7 bounding box candidates in each image after duplicate removal.

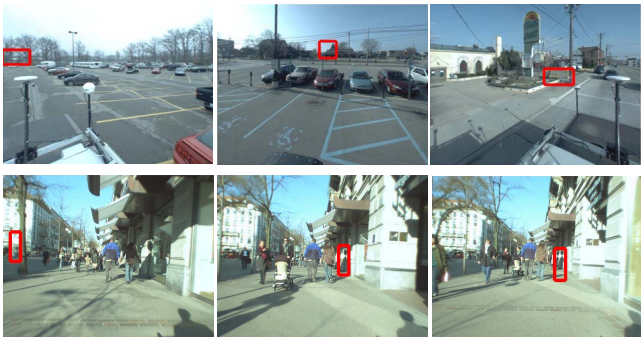

Figure 2. Incorrect bounding boxes on Ford car and Pedestrian datasets removed by our method. On each dataset, we rank the bounding boxes via the score $\frac{1}{K}\sum_{k=1}^{K}\|E_j^k\|_2$, where E_j^k denotes the j-th column of residue matrix E^k, and we pick the top three bounding boxes as examples here.

Table 2 shows the per-category performances of different methods in comparison, where we also directly cite the results from [10, 13, 31] for comparison. From the results, we observe the following: (1) Our proposed MLRR algorithm outperforms all the other baseline methods by a reasonable margin, achieving the best mean performance. (2) Both co-detection methods (MSR and MLRR) outperform the single-object detection methods (DPM, ESVMs and MFM). This intuitively is due to the fact that the former takes advantage of the consistent object appearance across the images while the latter only looks at one image at a time. (3) Our proposed MLRR performs significantly better than the other co-detection method MSR as it uses the low-rank constraint to capture the global structure of objects. On the other hand, MSR measures the mutual dependency of dif-

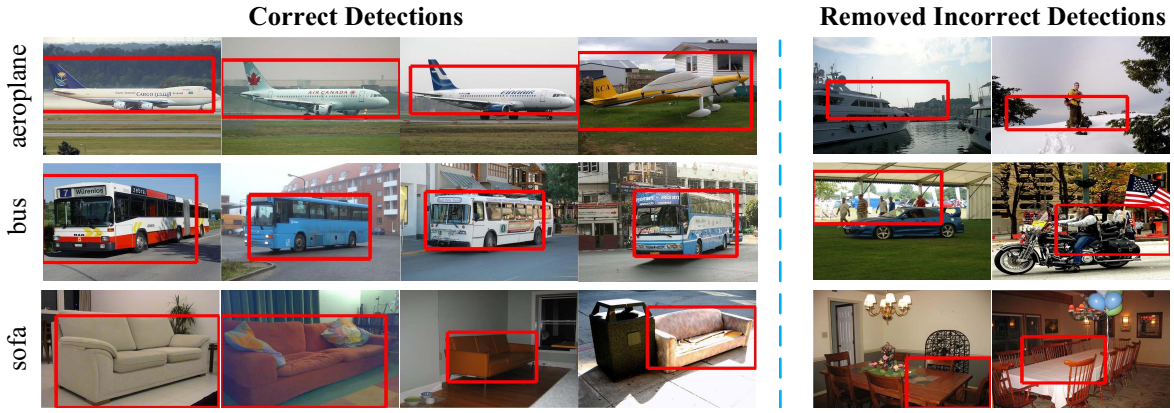

Figure 3. Example detection results and removed incorrect bounding boxes on PASVAL VOC 2007 test set. Red bounding boxes denote detections. Incorrect bounding boxes are picked from the top two bounding boxes ranked by scores $\frac{1}{K}\sum_{k=1}^{K}\|E_j^k\|_2$.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
DPM [12]	28.7	51.0	6.0	14.5	26.5	39.7	50.2	16.3	16.5	16.6	24.5	5.0	45.2	38.3	36.2	9.0	17.4	22.8	34.1	38.4	26.8
ESVMs [18]	20.8	48.0	7.7	14.3	13.1	39.7	41.1	5.2	11.6	18.6	11.1	3.1	44.7	39.4	16.9	11.2	22.6	17.0	36.9	30.0	22.7
Zhu *et al.* [31]	29.4	55.8	9.4	14.3	28.6	**44.0**	51.3	21.3	20.0	19.3	25.2	12.5	50.4	38.4	36.6	15.1	19.7	25.1	36.8	39.3	29.6
Desai *et al.* [10]	28.8	**56.2**	3.2	14.2	29.4	38.7	48.7	12.4	16.0	17.7	24.0	11.7	45.0	39.4	35.5	15.2	16.1	20.1	34.2	35.4	27.1
Harzallah *et al.* [13]	**35.1**	45.6	10.9	12.0	23.2	42.1	50.9	19.0	18.0	**31.5**	17.2	**17.6**	49.6	**43.1**	21.0	18.9	**27.3**	24.7	29.9	39.7	28.9
MFM	21.1	49.9	6.6	7.9	15.6	34.9	49.6	17.3	19.0	17.8	22.5	8.8	48.6	40.1	36.6	10.2	11.9	26.5	39.6	36.7	26.1
MSR	26.6	50.3	11.9	12.8	28.7	35.0	50.1	18.3	19.2	18.1	28.5	9.1	50.2	39.9	37.4	13.7	21.5	30.4	39.7	37.3	28.9
Our SLRR(SIFT)	33.0	52.2	11.6	16.1	29.0	42.5	52.0	19.2	19.2	22.7	29.4	9.6	50.6	41.0	38.5	17.1	24.1	31.3	41.1	39.7	31.0
Our SLRR(Gabor)	31.3	51.2	11.9	16.5	30.4	41.2	50.6	18.1	19.6	22.5	27.8	9.3	49.7	40.2	38.4	18.1	23.2	30.9	40.0	37.7	30.4
Our SLRR(LBP)	30.7	51.6	11.9	14.1	27.9	34.1	50.8	18.2	18.5	19.4	28.4	9.3	47.1	40.5	38.4	18.9	23.2	30.7	39.4	37.2	29.5
Our MLRR	34.1	53.0	**12.4**	**18.9**	**31.2**	43.2	**52.7**	**21.6**	**22.8**	25.0	**32.2**	10.6	**51.7**	41.0	**38.6**	**19.2**	**27.3**	**32.5**	**41.3**	**41.9**	**32.5**

Table 2. Performance comparison (AP %) on PASCAL VOC 2007 test set.

ferent objects separately and thus fails to leverage global information. (4) MLRR outperforms the performances obtained with only a single feature. This is because MLRR infers a shared low-rank coefficient matrix and can aggregate evidences from multiple features, resulting in a more cohesive representation. In Figure 3, we show some detection results and removed noisy bounding boxes by our method.

PASCAL VOC 2009 dataset. For VOC 2009, the annotations of the test samples are still confidential. Therefore, we use only the VOC 2009 "train/val" dataset which has $7,054$ images where the "train" set has 3473 images for training and the "val" set has $3,581$ images for test. For parameter selection, we use three-fold cross validation on the training set to determine the best parameter value for each method. The average recall rate across the 20 categories in the initial bounding box candidate pool is 61.2% and there are in average 6.1 bounding box candidates in each test image after duplicate removal.

In Figure 4, we show several precision-recall curves for example target categories. We can see that the precision of our method is higher than other methods as the recall varies. As seen on Table 3, our method outperforms all baselines and achieves the best performance on each category.

5.3. Discussion

Scalability. The scalability of our method is dictated by the size of the bounding box candidate pool during the co-detection process. Though large-scale co-detection is not the primal focus of this current work, we note that there are ways for controlling the complexity by dividing the test bounding boxes into clusters with moderate size and applying our method within each cluster.

Out-of-Sample Extension. For a new image, it is possible to apply the traditional out-of-sample extensions from transductive learning [3] to acquire detection scores of bounding boxes. When testing a new image, we begin by applying DPM and ESVMs on it to obtain its bounding box candidates as before. For each candidate z, we can use its low-level feature to search a set of nearest neighbors $\{x_i\}_{i=1}^{T}$ from all the bounding boxes in the original dataset, where x_i is a neighbor of z and T is the total number of the neighbors. Based on this set of neighbors, the detection score can be estimated as $s(z) = \sum_{i=1}^{T}\frac{W(z,x_i)}{\sum_{i=1}^{T}W(z,x_i)}s_i^*$, where $W(z,x_i)$ is the similarity between z and x_i, and s_i^* is the detection score of x_i obtained by our method. The result is a detection score for an unseen bounding box.

6. Conclusion

We have presented a robust object co-detection method to simultaneously detect target objects from an image collection. Given a bounding box pool represented in multiple feature spaces, we perform multiple linear reconstructions, each of which produces a reconstruction coefficient matrix measuring the mutual dependency of the bounding boxes. The co-detection problem is formulated as inferring a shared low-rank coefficient matrix across all reconstructions with noise and outlier removing constraints within each

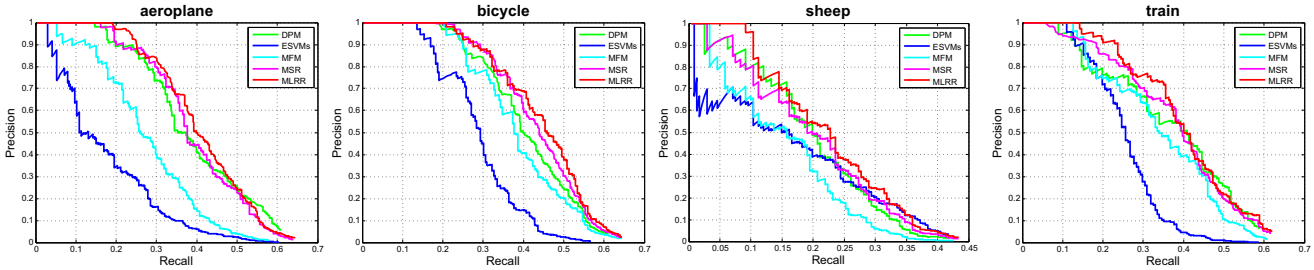

Figure 4. Precision-recall curves for four example categories from PASCAL VOC 2009 validation set.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
DPM [12]	40.3	45.7	10.3	10.1	23.6	47.1	36.8	17.9	16.3	14.7	13.7	11.5	41.1	32.9	44.2	9.9	18.5	19.2	38.2	26.2	25.9
ESVMs [18]	20.4	30.9	9.5	9.5	10.6	38.5	17.3	2.0	5.2	9.6	2.5	2.1	20.0	21.8	10.4	5.0	11.9	9.5	27.8	21.5	14.3
MFM	32.4	38.4	8.6	9.5	23.9	45.3	33.6	20.6	14.9	15.6	10.2	11.3	39.3	31.5	42.6	9.3	12.5	19.9	36.9	27.1	24.2
MSR	39.3	43.1	11.0	10.2	24.6	48.0	36.4	20.9	15.9	17.4	12.4	13.6	40.0	33.5	43.4	10.7	20.9	22.2	38.2	27.4	26.5
Our SLRR(SIFT)	42.6	48.4	12.5	11.7	28.3	49.3	39.3	24.0	17.2	19.6	15.9	13.9	43.3	34.3	44.5	13.0	23.7	23.5	41.6	28.2	28.7
Our SLRR(Gabor)	41.4	47.4	10.6	11.4	29.6	48.0	39.1	23.3	18.5	18.3	13.0	13.8	43.2	34.2	44.7	11.8	22.0	22.0	37.2	28.6	27.9
Our SLRR(LBP)	41.4	45.6	10.3	10.6	29.0	48.3	38.7	21.2	16.2	19.6	12.8	14.1	39.6	34.0	44.4	10.0	21.3	23.2	38.6	26.9	27.3
Our MLRR	**43.4**	**49.7**	**13.9**	**12.1**	**29.8**	**50.5**	**39.8**	**26.7**	**19.3**	**21.0**	**18.1**	**15.1**	**46.6**	**35.3**	**44.8**	**13.1**	**24.5**	**25.3**	**42.0**	**29.2**	**30.1**

Table 3. Performance comparison (AP %) on PASCAL VOC 2009 validation set.

feature. The low-rank coefficient matrix captures the global structure of objects across these multiple features and can be used to produce the co-detections using spectral clustering. Empirical experiment results on various object detection benchmarks show that our method outperforms the state-of-the-art generic object detection methods. For future work, we will investigate inductive object co-detection methods which not only infers a reconstruction coefficient matrix to leverage global structure but also builds a decision function for bounding boxes unseen in the candidate pool.

7. Acknowledgment

Xin Guo was supported by Chinese National Natural Science Foundation (90920001, 61101212), and National High and Key Technology R&D Program (2012AA012505, 2012BAH63F00), and National S&T Major Project of the Ministry of S&T 2012ZX03005008. Brendan Jou was supported by the U.S. National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *TPAMI*, 2012.
[2] S. Bao, Y. Xiang, and S. Savarese. Object co-detection. In *ECCV*, 2012.
[3] Y. Bengio et al. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. In *NIPS*, 2003.
[4] J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 2010.
[5] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 2011.
[6] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 2009.
[7] Y. Chen, L. L. Zhu, and A. Yuille. Active mask hierarchies for object detection. In *ECCV*, 2010.
[8] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan. Multi-task low-rank affinity pursuit for image segmentation. In *ICCV*, 2011.
[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
[10] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.
[11] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007.
[12] Felzenszwalb et al. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.
[13] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009.
[14] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.
[15] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.
[16] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *UAI*, 2009.
[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
[18] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.
[19] B. S. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *TPAMI*, 1996.
[20] L. Mukherjee et al. Analyzing the subspace structure of related images: Concurrent segmentation of image sets.
[21] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 2002.
[22] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 2000.
[23] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - Incorporating a global constraint into MRFs. In *CVPR*, 2011.
[24] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, 2012.
[25] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.
[26] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2011.
[27] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 2004.
[28] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, 2009.
[29] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, 2009.
[30] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In *CVPR*, 2010.
[31] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010.