

Individualness and Determinantal Point Processes for Pedestrian Detection

Donghoon Lee¹, Geonho Cha¹, Ming-Hsuan Yang², and Songhwai Oh¹

¹Electrical and Computer Engineering, Seoul National University, Korea

²Electrical Engineering and Computer Science, University of California at Merced

Abstract. In this paper, we introduce individualness of detection candidates as a complement to objectness for pedestrian detection. The individualness assigns a single detection for each object out of raw detection candidates given by either object proposals or sliding windows. We show that conventional approaches, such as non-maximum suppression, are sub-optimal since they suppress nearby detections using only detection scores. We use a determinantal point process combined with the individualness to optimally select final detections. It models each detection using its quality and similarity to other detections based on the individualness. Then, detections with high detection scores and low correlations are selected by measuring their probability using a determinant of a matrix, which is composed of quality terms on the diagonal entries and similarities on the off-diagonal entries. For concreteness, we focus on the pedestrian detection problem as it is one of the most challenging problems due to frequent occlusions and unpredictable human motions. Experimental results demonstrate that the proposed algorithm works favorably against existing methods, including non-maximal suppression and a quadratic unconstrained binary optimization based method.

Keywords: Determinantal point process, Individualness, Object detection, Pedestrian detection

1 Introduction

The goal of object detection is to locate objects from one known category in an image. It is essential for a number of vision tasks, such as visual tracking, scene understanding, and action recognition, to name a few. In visual tracking, tracking-by-detection is an effective approach which locates target objects in an image sequence by associating detections [1]. By learning about object locations in the image, we can better understand what is happening in the scene [2]. Object detection is also applied to action recognition by finding specific items related to an action of interest [3].

The general framework for object detection is to test image patches given by either a sliding window method or object proposals using a trained classifier. A number of object detectors have been developed which can reliably detect objects if they are well separated [4–7]. However, while it is desirable to report a single detection for each object, a sliding window method and object proposals entail a large number of raw detection responses around a true object as shown in Figure 1(a). Redundant detection responses are usually suppressed using a

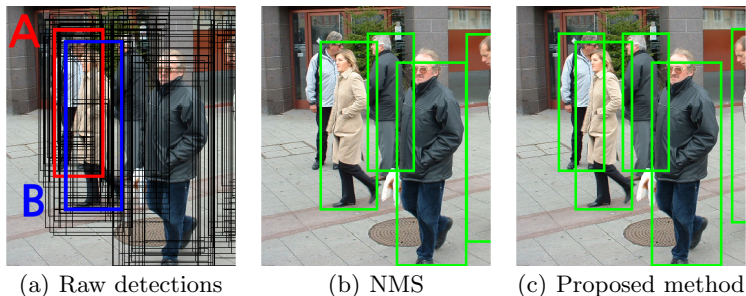


Fig. 1. Pedestrian detection results from the INRIA dataset. (a) Raw detections in black boxes using a sliding window method. (b) Detection results from a typical non-maximum suppression method. (c) Detection results from the proposed algorithm. Raw detection boxes A and B represent different pedestrians.

simple greedy algorithm, such as non-maximum suppression (NMS), based on the detection scores as shown in Figure 1(b).

It is known to be difficult to detect heavily occluded objects using the above-mentioned framework. This can be attributed to the fact that a detector is trained to distinguish between target classes, and not designed to differentiate between intra-class objects. For example, a detector reports detections A and B as shown in Figure 1(a). Since the detected bounding boxes are heavily overlapped, we examine whether these are false positives or negatives. During NMS, one of the detections, e.g., A, is likely to be suppressed, based on the confidence scores of these two detections. Therefore, the NMS scheme inevitably generates false negatives when multiple detections occur in proximity (i.e., in crowded scenes). On the other hand, if some prior information about two detections is available, e.g., detected objects have different identities, this issue can be alleviated. In fact, the false rejection significantly affects the detection accuracy, as the recall of raw detections in the dataset used in this paper is about 90%, while it drops to about 50% after final detections are selected by NMS. Therefore, it is an important task to select raw detections properly.

In this paper, we propose an algorithm based on *individualness* and a *determinantal point process* (DPP) for accurate detection which can be applied to any object detector. The proposed individualness is complementary to objectness for detection. While objectness is used to generate a set of detection candidates, individualness finds the relationship between candidates to obtain the final detection. We define individualness using the correlation between feature vectors, which describes the appearance and spatial information of detection results enclosed by bounding boxes. For concreteness, we focus on detecting multiple pedestrians in crowded scenes as it is one of the most challenging problems with numerous applications.

A DPP is a random process used to model particles with repulsive interactions in theoretical quantum physics [8]. It prohibits co-occurrences of highly correlated quantum states. This property is well suited for the task in this work to reject redundant detections. To apply a DPP, the quality and diversity factors need to be defined. They are used to compute the unary score (quality) and the

pair-wise correlation (diversity) of detections. Based on these two factors, we can select an optimal subset of detections as shown in Figure 1(c).

The contributions of this paper are summarized as follows. First, the problem in the existing detection framework, which naively selects final detections from a pool of detection candidates, is addressed. Second, a DPP is introduced to enhance the detection accuracy with a novel design of the quality term and the diversity feature. Finally, while finding the optimal subset of detections using a DPP is NP-hard, we show that a simple and efficient greedy algorithm performs favorably against existing pedestrian detection methods on the INRIA [5], PETS 2009 [9], and EPFL Terrace [10] datasets. On the PETS 2009 dataset with a deformable part model (DPM) detector, the proposed method achieves the accuracy rate of 41.9% and precision rate of 99.0%, while NMS achieves 23.2% accuracy and 98.2% precision. In addition, it takes less than 30 *ms* to process an image containing more than 300 detection candidates from over 30 pedestrians.

2 Related Work

Pedestrian Detection Methods. For completeness, we briefly discuss pedestrian detectors used in this paper. The HOG [5] features are computed by dividing an image patch into cells and blocks. Each block consists of cells and is represented by a histogram of gradients. Histograms from all blocks are concatenated into a single feature vector. Based on HOG features, an SVM classifier is trained and used to classify each sliding window in a test image for pedestrian detection. A deformable part model (DPM) [6] is developed based on HOG features. It learns a classifier for each body part and finds a pedestrian by considering body part detection scores and their spatial configuration. The DPM achieves higher accuracy in pedestrian detection but requires heavier computational load than the HOG based method.

A detector is developed based on different color spaces and gradients [11] where subregions or aggregated pixels of different channels are used to train a boosting-based object detector. This multi-channel based detector has been shown to perform more efficiently and accurately than HOG-based approaches. The faster R-CNN [12] method has been developed recently. This deep learning method is developed based on object proposals rather than sliding windows [5, 6, 11].

Merging Detection Results. We note that all the above-mentioned methods use NMS schemes to eliminate multiple detections around a true object including the state-of-the-art detector [12]. Given a set of detected bounding boxes within a region, NMS finds the one with the highest score and discards all the other neighboring detections. Typically, a neighboring detection is defined by thresholding on the overlap area ratio of bounding boxes. As stated in Section 1, NMS is likely to generate false positives or negatives when objects are close to each other.

Recently, numerous approaches have been proposed in order to address the discussed problems of NMS. In [13], it is shown that the localization accuracy of a bounding box is not strongly correlated to the score of a detection box when the score is high. To address this problem, a regression model is constructed to learn location statistics of raw detections with respect to the ground truth bounding

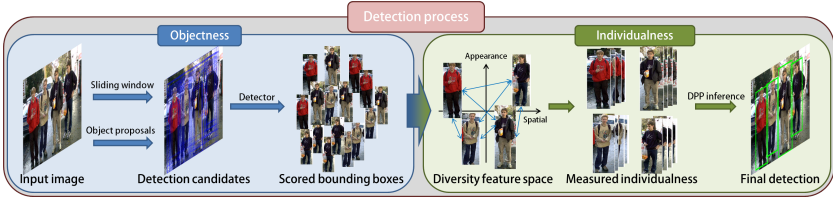


Fig. 2. Flow chart of the proposed detection framework.

boxes. However, a regression model needs to be trained for each detector. In addition, the performance of this method depends on the characteristics of the training data. The NMS scheme can also be integrated into deep learning models [14]. However, the parameters of NMS remain fixed during the training.

The task of NMS can be formulated as an optimization problem. In [15], a quadratic unconstrained binary optimization (QUBO) algorithm is proposed to replace NMS. The objective of QUBO is to find a binary vector where each element indicates whether the corresponding detection should be suppressed from a pool of candidate detections. The objective function consists of unary and pairwise terms. The unary term measures the confidence that a candidate detection truly represents a pedestrian and the pairwise term penalizes an overlap between a pair of candidate detections. The objective function is solved using a greedy algorithm. The main drawback of QUBO is that it approximates the distribution of raw detections, i.e., the distance between pedestrians, using a quadratic objective function. In Section 3.5, we show the limitations of QUBO in real world scenarios with comparisons to the proposed method.

A method based on affinity propagation clustering (APC) [16] is proposed. APC is a clustering algorithm which aims to find exemplars, such that the sum of similarities between exemplars and cluster members is maximized. However, APC is not naturally suitable for object detection since it does not explicitly penalize close exemplars which can be duplicated detections from the same object. Although this method can be improved by adding a repelling function which penalizes close exemplars, the detection accuracy is not notably increased compared to NMS.

3 Proposed Algorithm

The proposed detection process is a two-stage framework, which consists of objectness and individualness parts as shown in Figure 2. The objectness part returns a set of detection candidates with their scores. Then, the individualness part analyzes candidates using their appearances, relative locations, and sizes. The final detection is obtained by merging objectness (detection score) and individualness (similarity between detection candidates) into a single objective function. In this paper, we propose to use a DPP to model their relationships.

3.1 Determinantal Point Process Formulation

Let N be the number of items and $\mathcal{Y} = \{1, 2, \dots, N\}$ be the corresponding index set. Each item i is represented by its quality q_i and similarity S_{ij} to another

item j . A DPP aims to select high quality items while avoiding highly correlated items simultaneously. Let $Y \subset \mathcal{Y}$ be an index set of the selected items. Using their qualities and similarities, we compute a positive-semidefinite kernel matrix $L_Y = [L_{ij}]_{i,j \in Y}$, where

$$L_{ij} = q_i q_j S_{ij}. \quad (1)$$

To ensure the positive-semidefinite property, S_{ij} is usually computed by an inner product of each item's descriptor vector, which is called a diversity feature. Then, a DPP measures the probability, $\mathcal{P}_L(Y)$, of the selected indices using the determinant of L_Y . In other words, a DPP seeks to find the most probable subset by solving the following optimization problem:

$$Y^* = \arg \max_{Y \subset \mathcal{Y}} (\det(L_Y)) = \arg \max_{Y \subset \mathcal{Y}} \left(\prod_{i \in Y} q_i^2 \right) \det(S_Y), \quad (2)$$

where $S_Y = [S_{ij}]_{i,j \in Y}$ and Y^* is the optimal subset of item indices. Generally, the problem is NP-hard because all possible subsets have to be examined [17]. Fortunately, the problem is log-submodular which can be well approximated by a simple greedy algorithm [8].

The implicit meaning of the determinantal probability measure can be explained by the following example. With a subset $Y = \{i, j\}$,

$$\mathcal{P}_L(Y) \propto \begin{vmatrix} L_{ii} & L_{ij} \\ L_{ji} & L_{jj} \end{vmatrix} = \begin{vmatrix} q_i^2 & q_i q_j S_{ij} \\ q_i q_j S_{ij} & q_j^2 \end{vmatrix}. \quad (3)$$

The diagonal entries are computed without a similarity term because the correlation to itself is always one. The determinant is decreased when $|S_{ij}|$ increases. Therefore, a DPP tends to pick uncorrelated items in these cases. On the other hand, higher quality items increase the determinant, and thus a DPP tends to pick high-quality items in such cases. As a consequence, an appropriate design of the quality term and the diversity feature plays an important role. In this work, we use a DPP for the detection problem by using each raw detection as an item and Y^* as the final detection set.

3.2 Quality Term

The quality term indicates the value of an item. For the pedestrian detection problem, it can be described by a detection score. However, the original scores are independently obtained from each image patch. Therefore, they do not contain information of neighboring detections or the scene. We propose a simple, yet effective scheme to re-score each detection.

Let $\mathbf{s}^o = \{s_1^o, s_2^o, \dots, s_N^o\}$ be set of scores for N raw detections. One common factor that degrades the detection accuracy is a wrong detection with a high confidence score, as it can potentially suppress neighboring true detections. The problem is worsened when the bounding box of the wrong detection is large. Figure 3(a) shows an example of wrong detections from a detector. In order to deal with this problem, we penalize unnecessarily large raw detections. As such, we count the number of other raw detections inside a bounding box. Figure 3(b) shows the number of raw detections inside a ground truth bounding box from the INRIA dataset. It shows that a bounding box with a small number of

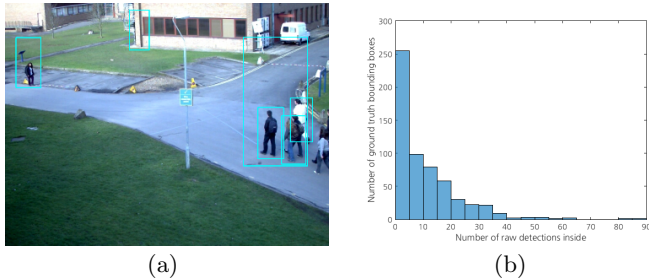


Fig. 3. (a) Detection results show that a detector can return wrong detections with large bounding boxes. These wrong detections may suppress other true detections. (b) The number of raw detections inside a ground truth bounding box (INRIA dataset).

raw detections is more likely to contain ground truth detections. Based on this observation, we re-score each detection as $s_i^c = s_i^o \exp(-\lambda n_i)$, where n_i is the number of raw detections inside the current bounding box and λ is a constant. Another advantage of the proposed re-scoring function is that it favorably yields tight bounding boxes as discussed in Section 4.

Additionally, we can use prior information when it is available. For a fixed camera environment, such as the PETS 2009 and EPFL Terrace datasets, and numerous practical surveillance applications, the height of a pedestrian at each location in an image does not vary significantly. Let (x_i, y_i) be the image coordinate location and h_i be the height of detection i , which is the height of its bounding box. Given a training set, we can find coefficients of the following relationship to estimate the expected height of a person at different locations:

$$\tilde{h}_i = ax_i + by_i + c. \quad (4)$$

Then, we re-score each detection based on the assumption that the height distribution of people is Gaussian, i.e., $s_i^c = s_i^o \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(h_i - \tilde{h}_i)^2}{2\sigma^2}\right)$. The re-scored detection score, $\mathbf{s} = \{s_1, s_2, \dots, s_N\}$, is obtained by using s_i^p when prior information is available, otherwise s_i^c . The quality term $\mathbf{q}$ is represented as follows:

$$\mathbf{q} = \alpha \mathbf{s} + \beta, \quad (5)$$

where α and β are weights for the quality term. The weights are needed to balance the detection score of different detectors, e.g., an average detection score of a DPM detector is about 0.7 while an ACF detector is about 33.2 from the INRIA dataset. We find these parameters using the pattern search algorithm [18], which maximizes the detection performance on a training set.

3.3 Individualness and Diversity Feature

Individualness aims to determine whether two image patches originated from the same object (person) or not. It might be reminiscent of the person re-identification problem in a multi-camera environment. However, this problem is

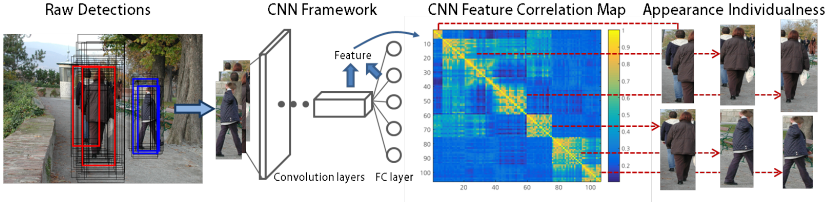


Fig. 4. Measuring the appearance individualness of each detection box using a CNN. The goal is to give a high correlation score to the bounding boxes around a single person (blue boxes), while giving a low correlation score when there is a different person.

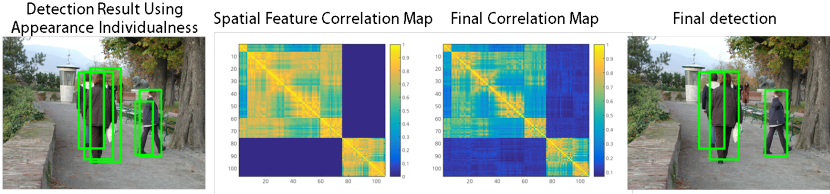


Fig. 5. The result of DPP inference. It is more robust to detect pedestrians by combining appearance and spatial information.

different and difficult for two reasons. First, the overlapping region of two image patches contains exactly the same information. Therefore, the distance between two image patches in a feature space tends to be closer. Second, we mainly deal with occluded people while the person re-identification problem typically takes image patches of well-separated people as inputs.

We tackle this problem by measuring the correlation of feature descriptors from bounding boxes. The feature descriptor should be insensitive to new background pixels and scale variations around a single person as shown in blue detection boxes in Figure 4, while being sensitive to a new pedestrian as shown in red detection boxes in Figure 4. Toward this goal, we consider features from convolutional neural network (CNN) layers. The CNN features are translation and scale invariant due to multiple max-pooling operations. Figure 4 shows the correlation between CNN features of different bounding boxes using a pre-trained network for image classification [12]. Overall, the correlation matrix is block diagonal and the correlation between different individuals is low. This observation encourages us to use the CNN feature, denoted as ϕ_i for the i -th detection, for the individualness.

While highly effective, solely use the CNN features to measure the individualness is not robust. As shown in Figure 4, there can be duplicated clusters for a single pedestrian or ambiguous clusters between nearby pedestrians. Consequently, there may be a wrong inference from the DPP algorithm as shown in Figure 5. In this example, the shapes inside two boxes around a boy on the right-hand side are significantly different although a large portion of the smaller bounding box is included in the larger box. The difference misleads the CNN feature to have a low correlation between these two boxes (see the CNN feature correlation map in Figure 4). On the other hand, in the case of the woman in

the middle, the CNN features fail to ignore new background pixels. To overcome this problem, we additionally consider the spatial location of each detection box.

The spatial individualness is designed to give high correlation to multiple detection boxes around a single pedestrian. Let φ_i be the spatial individualness vector of the i -th detection and let π_i denote a set of pixel indices belonging to the detection box. We propose an efficient form as follows:

$$\varphi_i^k = \frac{1}{\sqrt{|\pi_i|}} \begin{cases} 1 & \text{if } k \in \pi_i \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where φ_i^k is the k -th entry of $\varphi_i \in [0, 1]^n$, n is the number of pixels in an input image, and $|\pi_i|$ is the number of pixels in the i -th detection box. The square root of $|\pi_i|$ is used for normalization so that φ_i is unit norm. Although the dimension of φ_i is as large as the size of the image, $\varphi_i^\top \varphi_j$ can be computed easily as follows:

$$\varphi_i^k \varphi_j^k = \frac{1}{\sqrt{|\pi_i||\pi_j|}} \begin{cases} 1 & \text{if } k \in \pi_i \cap \pi_j \\ 0 & \text{otherwise} \end{cases}. \quad (7)$$

Furthermore, φ_i itself does not have to be stored in the memory since only detection sizes and the overlap area are required

$$\varphi_i^\top \varphi_j = \frac{|\pi_i \cap \pi_j|}{\sqrt{|\pi_i||\pi_j|}} \in [0, 1]. \quad (8)$$

This term is designed to increase the correlation when there is more overlap. It also satisfies zero correlation for non-overlapping detections and the correlation with itself is one.

Given individualness features ϕ_i and φ_i , there are several schemes to merge them into a single diversity feature. For example, a structured DPP [19] is performed by averaging the values of different feature descriptors. However, this method is restrictive because the number of features must be the same. We propose a more general and effective way to design a diversity feature and construct a positive semi-definite similarity matrix S directly. Let S^c and S^s denote the similarity matrices constructed from ϕ_i and φ_i , respectively. In other words, $S_{ij}^c = \phi_i^\top \phi_j$ and $S_{ij}^s = \varphi_i^\top \varphi_j$. We then merge them into a single matrix using an operation that preserves the positive semi-definite property.

In this paper, two different ways are considered. First, from the Schur-product theorem [20], $S = S^c \circ S^s$ is a positive semi-definite matrix where $\circ$ is a Hadamard product or an element-wise product of matrices. For pedestrian detection, we find this approach is not suitable, as the correlation between items becomes too small when multiplying values within $[0, 1]$. Second, any positive combination of positive semi-definite matrices is also positive semi-definite. Therefore, we construct S as follows:

$$S = wS^c + (1 - w)S^s, \quad (9)$$

where $0 \leq w \leq 1$ can determine the relative importance of each feature descriptor. This formulation implies that the diversity feature is a concatenation of ϕ_i and φ_i . We set w as 0.8 throughout the paper in order to make sure that the correlation between items that are spatially separated is low enough. In the next section, we discuss how to efficiently solve (2) given the proposed quality term and similarity matrix.

Algorithm 1 A greedy algorithm for solving (2)

```

1: INPUT :  $\mathbf{q}, S, \mathcal{Y}, \epsilon$ 
2: OUTPUT :  $Y^*$ 
3:  $Y^* = \emptyset$ 
4: while  $\mathcal{Y} \neq \emptyset$  do
5:    $j^* = \arg \max_{j \in \mathcal{Y}} (\prod_{i \in Y^* \cup \{j\}} q_i^2) \det(S_{Y^* \cup \{j\}})$ 
6:    $Y = Y^* \cup \{j^*\}$ 
7:   if  $P_L(Y)/P_L(Y^*) > 1 + \epsilon$  then
8:      $Y^* \leftarrow Y$ 
9:     delete  $j^*$  from  $\mathcal{Y}$ 
10:  else
11:    break
12:  end if
13: end while

```

3.4 Mode Finding

As mentioned earlier, the problem of finding the exact solution to (2) is NP-hard [17]. Since $\mathcal{P}_L(Y)$ is log-submodular, greedy mode finding approaches for DPPs perform well in numerous machine learning applications [8]. Using a similar idea, our algorithm iteratively adds the j^* -th detection to the final solution set Y^* if j^* maximizes $\mathcal{P}_L(Y^* \cup \{j^*\})$ among the remaining detection candidates. Once j^* is added to the Y^* , we delete j^* from the candidate detection set $\mathcal{Y}$. The algorithm terminates when the candidate detection set is empty or there is no more detection which increases $\mathcal{P}_L(Y)$ by $(1 + \epsilon)$ times the previous value of $\mathcal{P}_L(Y)$. The main steps of the proposed algorithm are summarized in Algorithm 1. Note that although there exists an approximate algorithm [21] for this problem, we use Algorithm 1, which has a formal guarantee for monotone submodular problems [8], since it is fast and works well in practice.

3.5 Relationship to Quadratic Unconstrained Binary Optimization

The objective function of quadratic unconstrained binary optimization can be converted to a similar form using the DPP objective function as follows:

$$\max_x x^\top Lx = \max_Y \sum_{i,j} (L_Y)_{ij}, \quad (10)$$

where x is a binary vector and Y is a set of non-zero indices in x . In other words, QUBO finds Y that maximizes the sum of all elements in a submatrix L_Y , while a DPP seeks to maximize the determinant of L_Y . This leads to two key differences. First, QUBO cannot deal with positively correlated items. By (1), those items have positive entries in L_Y . Therefore, QUBO blindly selects them all to maximize the objective function. On the other hand, the DPP is well-defined for both positively and negatively correlated items. Second, QUBO penalizes highly correlated items more than DPP, which is not suitable to detect occluded pedestrians. We show this by an illustrative example that we face often during the experiments. Let there be two pedestrians in an image, and a detector reports a detection for each of them which constructs $L_Y = \begin{bmatrix} 2 & -0.8 \\ -0.8 & 1.4 \end{bmatrix}$. We set off-diagonal entries as -0.8 to represent overlapped detections. The first pedestrian has higher detection score, which usually indicates that the second

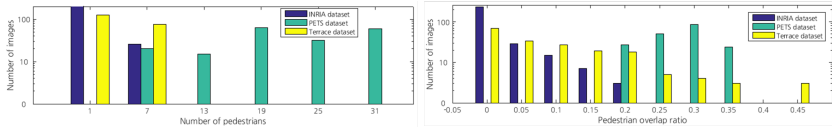


Fig. 6. Number of pedestrians and their overlap ratio for each evaluated dataset. The pedestrian overlap ratio of an image is defined by dividing the summation of overlapped bounding box regions by the summation of all bounding box areas. The numbers are calculated based on the ground truth data. Note that the y -axis is log-scale.

pedestrian is occluded by the first pedestrian. In this case, QUBO and DPP work differently. By Algorithm 1, the first pedestrian is picked (QUBO uses a similar greedy algorithm). Then, QUBO does not pick the second detection since $-0.8 - 0.8 + 1.4 = -0.2 < 0$ while a DPP selects both detections since $\det(L_Y) = 2.16 > 2$. Moreover, QUBO ignores the elements of the previously selected items while DPP considers the whole matrix to select a new item. For instance, if the first element of L_Y is 1.5, DPP does not pick the second item, since $\det(L_Y)$ becomes 1.46. The additional consideration enables a DPP to deal with more complex relationships between items. In Section 4, we demonstrate that the proposed method outperforms both NMS and QUBO for detecting pedestrians.

4 Experiments

We first discuss the experimental settings for evaluating the proposed and existing methods, and then present the empirical results. More results are available in the supplementary material. All the source code and annotated datasets will be made available to the public.

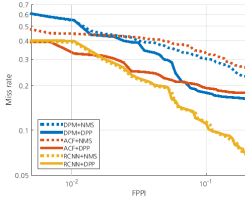
4.1 Experimental Settings

We evaluate the proposed algorithm with comparisons to other methods on the INRIA [5], PETS 2009 [9], and EPFL Terrace [10] datasets. Figure 6 shows the number of pedestrians per image and the average overlap ratio of pedestrians of these datasets, where the average overlap ratio is defined by $\frac{\sum_{i,j} |\pi_i \cap \pi_j|}{\sum_i |\pi_i|}$, and π_i is the same as in Section 3.3.

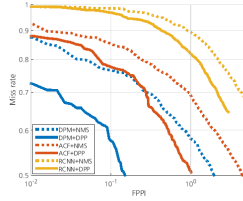
The INRIA dataset contains a relatively small number of well-separated pedestrians as it is designed to measure the effectiveness of a detector. We use 288 images in the test set for evaluation. For the PETS 2009 dataset, we use the walking sequence (S1.L1) of 190 frames and the dataset contains at most 33 people in a frame. This sequence results in a significant number of overlaps between people because the set is originally designed for the pedestrian density estimation. The height of a pedestrian on the image coordinate gradually decreases to half of the maximum height as a pedestrian moves from the lower right corner to the upper left corner of the image. To reliably detect small pedestrians at the upper left corner, we have resized each image from 768×576 pixels to 1440×1080 pixels in all experiments. We have randomly selected 50 frames

Table 1. Pedestrian detection results. We report results by setting a false positive per image (FPPI) to 0.1. (TP=number of true positives, FP=number of false positives, FN=number of false negatives, Accuracy=TP/(TP+FP+FN), and Precision=TP/(TP+FP).)

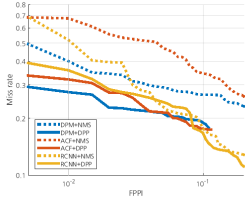
		DPM +NMS	DPM +DPP	ACF +NMS	ACF +DPP	RCNN +NMS	RCNN +DPP
INRIA	TP	409	481	391	475	521	523
	FP	29	29	29	29	29	29
	FN	180	108	198	114	68	66
	Accuracy	66.2	77.8	63.3	76.9	84.3	84.6
	Precision	93.4	94.3	93.1	94.2	94.7	94.7
PETS 2009	TP	1017	1836	633	967	69	137
	FP	19	19	19	19	19	19
	FN	3348	2529	3732	3398	4296	4228
	Accuracy	23.2	41.9	14.4	22.1	1.6	3.1
	Precision	98.2	99.0	97.1	98.1	78.4	87.8
EPFL Terrace	TP	525	579	467	590	580	594
	FP	20	20	20	20	20	20
	FN	192	138	250	127	137	123
	Accuracy	71.2	78.6	63.4	80.1	78.7	80.6
	Precision	96.3	96.7	95.9	96.7	96.7	96.7



(a) INRIA dataset



(b) PETS 2009 dataset



(c) EPFL Terrace dataset

Fig. 7. Detection error tradeoff curves for different detectors and datasets. The x -axis is false positive per image (FPPI). The lower the curve is better.

from another sequence with the same viewpoint for learning parameters. The EPFL Terrace dataset has 5,010 frames at a frame rate of 25 fps and there are at most seven people in a frame. We use every 25th frame from Sequence-1 of Camera-3, resulting in a total of 201 frames for evaluation. It is recorded from a relatively short distance, therefore, sometimes the height of a person is taller than the height of the image.

To measure detection performance, we compare the results from each evaluated algorithm to the ground truths. Let d_e be the detection reported by an algorithm and d_g be the ground truth detection. Each detection d_e is declared as a true positive when d_e satisfies the PASCAL 2012 detection criteria [22]. Since multiple detections on a single object should be false positives except one, we run the Hungarian algorithm using area ratios as costs to find the best matching between detections and ground truths.

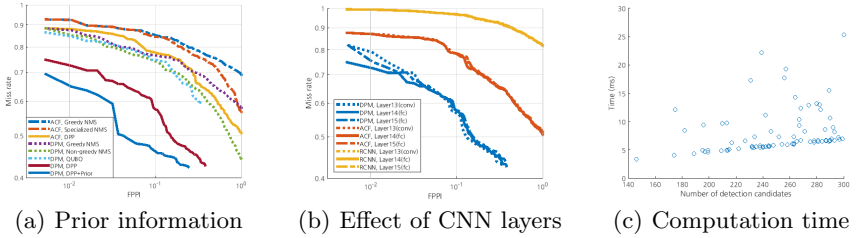


Fig. 8. (a) Comparison to greedy NMS, non-greedy NMS, and QUBO, including the effect of the prior information, on the PETS 2009 dataset. (b) The effect of different CNN features. (c) The computation time of the proposed algorithm.

4.2 Evaluation Results

The quantitative detection performance of the proposed algorithm is reported in Table 1 and Figure 7. It contains results of three detectors: DPM [6], ACF [11], and faster R-CNN [12], using the original source codes. For fair comparisons, we apply NMS using [22] to define neighboring detections for all detectors. The results show that the proposed algorithm performs favorably against NMS for all detectors and datasets. The improvement of the detection accuracy is most noticeable on the PETS dataset. This is because the problem of selecting a correct set of individual raw detections becomes more important when there are frequent occlusions due to crowded pedestrians. It is also interesting to see that the detection accuracy of the RCNN detector outperforms other detectors on the INRIA dataset while it is less effective on the PETS dataset. This can be explained by noting that the object proposal method tends to generate boxes on a group of overlapping pedestrians instead of boxes on individuals, even when a large number of proposals are used. Note that the proposed algorithm do not use any prior information for this experiment.

In addition to the basic greedy NMS, which is still used in many state-of-the-art detectors, we also report results from other methods in Figure 8(a). Instead of [22], the ACF detector often uses a different criteria, $\frac{\text{area}(d_g \cap d_e)}{\min(\text{area}(d_g), \text{area}(d_e))} > 0.65$, for NMS (denoted as the Specialized NMS). Compared to the criteria, the proposed algorithm generates more accurate results. The non-greedy NMS method examines all pairs of detections. In other words, a detection can suppress other detections after being suppressed, whereas in the greedy NMS, a detection can be eliminated before it can suppress other detections. Therefore, non-greedy NMS tends to give a small number of false positives while it is computationally expensive. Nevertheless, the results by the non-greedy NMS are worse than those by optimization based algorithms. The accuracy of QUBO is similar to non-greedy NMS which implies that the QUBO formulation is less effective for detecting pedestrians. On the other hand, the proposed algorithm performs better than other approaches and can be further improved using prior information when it is available.

The effects of different CNN layers are shown in Figure 8(b). We consider the 13-th layer (convolution layer), 14-th layer (fully-connected layer), and 15-th layer (fully-connected layer) features of the faster R-CNN to compute individual-

ness. We evaluate all combinations of detectors and layers. For example, DPM, Layer14(fc) in Figure 8(b) is the result of feeding raw detection boxes of the DPM to the faster RCNN and use the 14-th layer as a feature to compute individualness. The results show that the detection accuracy is not sensitive to the selected features from different layers. Feature combinations of two or more layers are not shown in the figure for clearer illustration since they achieve similar results. We use the 4,096-dimensional vector from the 14-th layer as the diversity feature ϕ .

The run time performance of the proposed algorithm is shown in Figure 8(c) with respect to the number of detection candidates in the scene. We use the faster R-CNN code which generates a maximum of 300 object proposals. The run time includes the execution of Algorithm 1, excluding the time spent by a detector. For detectors that are not based on the CNN architecture, an extra time of 248 *ms* is needed to compute the convolutional features (image patch resizing and feed forward) of 300 detection candidates on a machine with Intel Xeon 2.3GHz CPU, 128GB memory, and GeForce GTX Titan X D5 12GB GPU. For images that have the maximum number of object proposals, we report the average of the execution times. On average, it takes less than 30 *ms* on MATLAB, demonstrating the efficiency of the proposed algorithm.

The localization accuracy of a bounding box, $\frac{|d_e \cap d_g|}{|d_e|}$, is also measured. For the PETS 2009 dataset, it is 0.81 using the proposed algorithm, while it is 0.76 using NMS by averaging the results from all matched detections. It demonstrates that the proposed algorithm yields tighter bounding boxes. Figure 9 shows detection results by different algorithms on the PETS 2009 and EPFL Terrace datasets based on the DPM detector. The proposed algorithm generate accurate detection results and tighter bounding boxes. For example, at the 84-th frame of the EPFL Terrace dataset, both NMS and QUBO fail to detect an occluded pedestrian in the middle while the proposed algorithm returns correct detections.

5 Conclusions

We present an algorithm for improving detection performance by introducing individualness. Individualness measures the similarity between detection candidates while the objectness aims to generate the candidates with scores. The appearance and spatial information of each detection candidate are considered to compute individualness. Then, a determinantal point process combines the score and similarities to obtain final detections. Experimental results show that the proposed algorithm outperforms non-maximum suppression and QUBO. Furthermore, the proposed algorithm takes less than 30 *ms* to process an image with 300 detections from over 30 pedestrians.

Acknowledgements. The work of D. Lee, G. Cha, and S. Oh is supported in part by a grant to Bio-Mimetic Robot Research Center funded by Defense Acquisition Program Administration and Agency for Defense Development (UD130070ID) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2015R1A2A1A15052493). The work of M.-H. Yang is supported in part by the NSF CAREER grant #1149783, and gifts from Adobe and Nvidia.

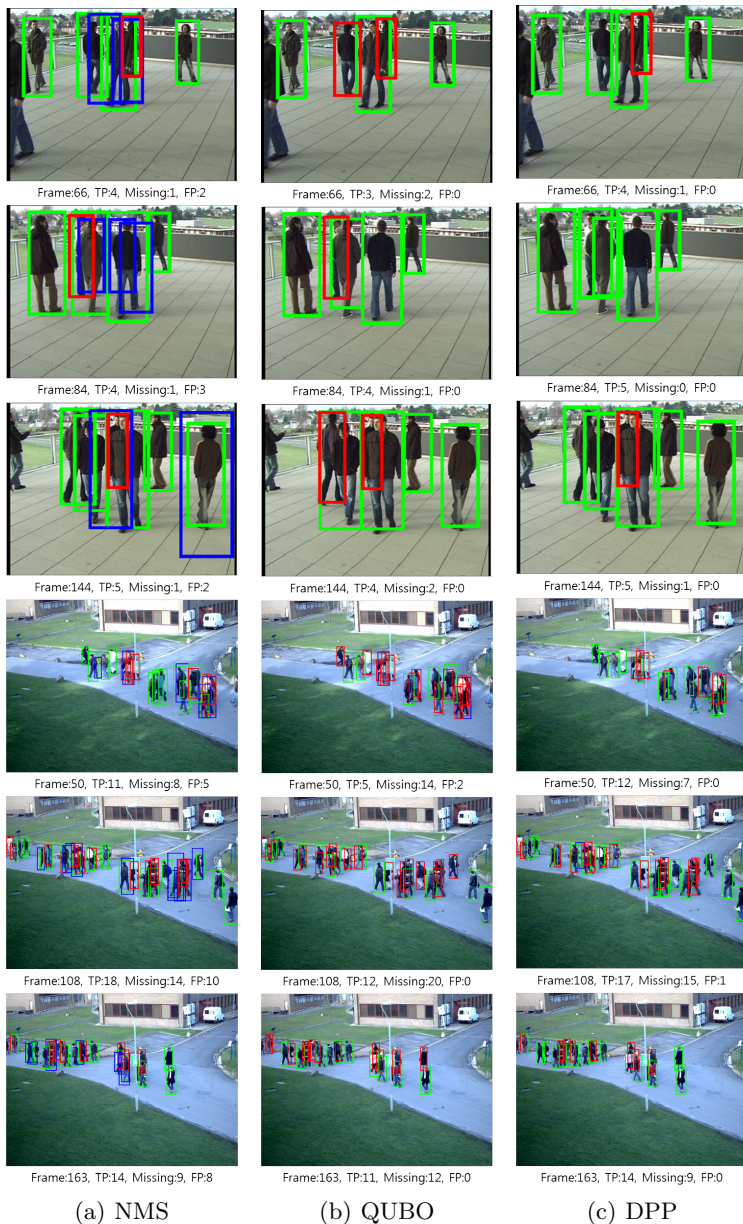


Fig. 9. Some detection results from the EPFL Terrace and PETS 2009 datasets. A green box is a true positive, a blue box is a false positive, and a red box is a missing detection. (a) Results from the DPM detector using NMS. (b) Results from the DPM detector using QUBO. (c) Results from the DPM detector using DPP.

References

1. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: Proc. of the IEEE Computer Vision and Pattern Recognition. (2008) [1](#)
2. Li, L.J., Socher, R., Fei-Fei, L.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: Proc. of the IEEE Computer Vision and Pattern Recognition. (2009) [1](#)
3. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: Proc. of the IEEE International Conference on Computer Vision. (2011) [1](#)
4. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. of the IEEE Computer Vision and Pattern Recognition. (2001) [1](#)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of the IEEE Computer Vision and Pattern Recognition. (2005) [1](#), [3](#), [10](#)
6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(9) (2010) 1627–1645 [1](#), [3](#), [12](#)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. of the IEEE Computer Vision and Pattern Recognition. (2014) [1](#)
8. Kulesza, A., Taskar, B.: Determinantal point processes for machine learning. arXiv preprint arXiv:1207.6083 (2012) [2](#), [5](#), [9](#)
9. Ferryman, J., Shahrokni, A.: PETS2009: Dataset and challenge. In: Winter-PETS. (2009) [3](#), [10](#)
10. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(2) (2008) 267–282 [3](#), [10](#)
11. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(8) (2014) 1532–1545 [3](#), [12](#)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. (2015) 91–99 [3](#), [7](#), [12](#)
13. Liu, S., Lu, C., Jia, J.: Box aggregation for proposal decimation: Last mile of object detection. In: Proc. of the IEEE International Conference on Computer Vision. (2015) 2569–2577 [3](#)
14. Wan, L., Eigen, D., Fergus, R.: End-to-end integration of a convolution network, deformable parts model and non-maximum suppression. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 851–859 [4](#)
15. Rujikietgumjorn, S., Collins, R.T.: Optimized pedestrian detection for multiple and occluded people. In: Proc. of the IEEE Computer Vision and Pattern Recognition. (2013) [4](#)
16. Rothe, R., Guillaumin, M., Gool, L.V.: Non-maximum suppression for object detection by passing messages between windows. In: Proc. of the IEEE Asian Conference on Computer Vision. (2014) [4](#)
17. Çivril, A., Magdon-Ismael, M.: On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science* **410**(47) (2009) 4801–4811 [5](#), [9](#)
18. Meinshausen, N., Bickel, P., Rice, J.: Efficient blind search: Optimal power of detection under computational cost constraints. *The Annals of Applied Statistics* (2009) 38–60 [6](#)
19. Kulesza, A., Taskar, B.: Structured determinantal point processes. In: Advances in neural information processing systems. (2010) 1171–1179 [8](#)

20. Schur, J.: Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und Angewandte Mathematik* **140** (1911) 1–28 [8](#)
21. Gillenwater, J., Kulesza, A., Taskar, B.: Near-optimal MAP inference for determinantal point processes. In: *Advances in Neural Information Processing Systems*. (2012) [9](#)
22. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge 2012 (VOC2012) results [11](#), [12](#)