People Detection in Crowded Scenes by Context-driven Label Propagation

Jingjing Liu
Rutgers University
jl1322@cs.rutgers.edu

Quanfu Fan
IBM T. J. Watson Research Center
qfan@us.ibm.com

Sharath Pankanti
IBM T. J. Watson Research Center
sharat@us.ibm.com

Dimitris N. Metaxas
Rutgers University
dnm@cs.rutgers.edu

Abstract

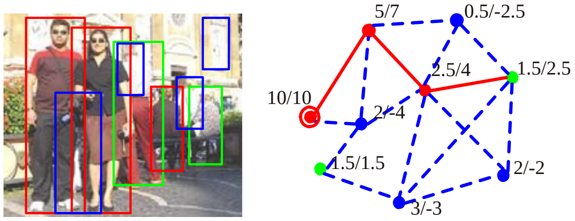
Exploiting contextual cues has been a key idea to improve people detection in crowded scenes. Along this line we present a novel context-driven approach to detect people in crowded scenes. Based on a context graph that incorporates both geometric and social contextual patterns in crowds, we apply label propagation to discover weak detections contextually compatible with true detections while suppressing irrelevant false alarms. Compared to previous approaches for context modeling limited to only pairwise spatial interactions between local object neighbors, our approach provides a more effective way to model people interactions in a global context. Our approach achieves performance comparable to state of the art on two challenging datasets for people and pedestrian detection.

1. Introduction

People detection in images is a fundamental vision problem, which is central to a wide range of applications such as video surveillance, robotics, and autonomous driving. The problem has been extensively studied in computer vision [3], and recent years have witnessed great advances in approaches such as Deformable Part Models [12], Poselets [4] and deep convolutional neural networks [16]. Although these approaches and their variants have achieved tremendous promising results, the problem still remains quite challenging when it comes to the scenario where cluttered background, various occlusions, and large pose variations exist.

Despite these difficulties, some research effort has exploited contextual cues in a scene to improve people detection in adverse conditions. For example, two-person or multiple-person classifiers were built directly in several approaches [37, 27] to handle partial occlusion. Other approaches [7, 32, 40] explored pairwise spatial relationships between local neighbors to boost detection performance, under the framework of structural learning.

Inspired by the above approaches, we develop a new framework to address the challenges for people detection by putting people in a global context and modeling their interactions. In a crowded scene, people usually form groups where they interact with each other, both geometrically and socially [29, 31, 40, 5]. A group of people, either queuing, sitting or walking together, indicate spatial closeness and similar scales (Fig. 1). There also exist strong social patterns in the group such as *facing* (i.e., two people face each other) and *following* (i.e., people stand or sit side by side).



(a) (b)

Figure 1. Context-drive label propagation for people detection. (a) Grouped people tend to present spatial closeness and similar scales (image from VOC 2012 [11] for illustration). There also exists social interactions in a group such as *facing* and *following* (e.g., the two people on the left stand side by side). (b) A context graph captures the interaction strength between human hypotheses (or detections). Each node here is a human hypothesis from an underlying detector. True detections are colored as either red (high confidence) or green (low confidence) while false alarms as blue. A bold edge indicates mutual *attraction* between two nodes, i.e. they are contextually compatible. Oppositely a dotted edge suggests an opposite relationship, i.e. *repulsion*. Our approach applies label propagation to boost up weak detections (greed nodes) while suppressing irrelevant false alarms.

To effectively leverage the geometric and social contexts in crowds, we propose a unified framework for people detection that integrates visual recognition with graph-based context modeling. We formulate the detection task as an optimization problem where the goal is to find a maximum set of human hypotheses that agrees on both visual detections and their contextual interactions in an image. While such optimization problem is theoretically intractable, we show that it can be approximately addressed by *label propagation* [45] in a progressive way. Label propagation was originally proposed for predicting unlabeled instances from labeled data [45]. It propagates labels to data iteratively by proximity. In our case, true detections are supposed to be contextually concordant with each other, but incompatible to false alarms. This suggests that strong detections with high confidence can boost up weak ones by spreading rewards through contextual proximity and meanwhile penalize false positives according to contextual incompatibility, in a similar spirit to label propagation.

More specifically, our approach starts by taking input from an underlying detector, possibly with a large number of false alarms. We build a context graph that incorporates contextual information available in a scene (spatial, scale, social, and overlap cues) for label propagation. For the purpose of rewarding true detections as well as suppressing irrelevant false alarms, our approach enables the graph to spread both positive and negative contextual potentials along edges in the context graph, which depends on contextual attraction or repulsion. As a result, contextually compatible human hypotheses get reinforced by receiving positive potentials during the propagation while false alarms are contained due to being negated by their incompatibility with true detections. The idea of our approach is illustrated in Fig. 1. Since we do not have true 'labeled' data to start with, we further design a greedy-like inference which iteratively adds the best hypothesis with the most potential gain at each time to initialize a new round of propagation. The process repeats until convergence when no new hypothesis can be instanced.

The main contribution of our approach is the application of graph-based label propagation to exploit contextual information for people detection. Compared to the structure learning framework modeling contextual interactions only between local neighbors [7, 31, 40], our approach is clearly advantageous in that the graph-based propagation make interactions between any two nodes possible. Such a capability allows our approach to discover challenging weak human instances, even if they are not close to any strong detection, as illustrated later in Fig. 5. We would also like to point out that Conditional Random Fields (CRFs) [20] methods, while being another option for contextual modeling, are less suitable for our problem. This is largely because capturing long-range interactions requires fully connected CRFs, which need expensive inference mechanisms or strong assumptions such as Gaussian edge potentials [19] that does not fit our problem. Besides, modeling of fully connected CRFs requires large-scale and fine annotations covering contextual interactions.

We validate our approach using two challenging crowd datasets, one for people detection with variations in pose and size, and the other for pedestrian detection in low-resolution images. The experimental results confirm that our method can significantly improve people detection in crowded scenarios, achieving performance comparable to state of the art reported in the literature.

2. Related Work

Over the past decades, significant progress has been made for people detection. A large body of pioneering work has focused on developing low-level image representations for human, for example, rectangular features [35], Histograms of Oriented Gradients (HoG) [6], Local Binary Pattern (LBP) [25], adaptive contour feature [14], spatially pooled features [28], etc. These approaches commonly assume people are almost upright (pedestrian) and hardly occluded. A nice review on pedestrian detection can be found in [9, 3]. In recent years, another trend that exploits part-based model arises from applications in the scenario where people expose variant poses and may be partial visible [39, 21, 1, 13, 41, 38, 42]. The most famous among those methods ought to be two techniques, namely Poselet [4] and Deformable Part Models (DPM) [12]. Rather than a single detector, DPM is consisted of a mixture of components (part-based models), combined at some deformation cost which enables the model to compromise intra-category variance. Poselet trains a set of HoG classifiers, based on diverse keypoint configurations.

Recently, deep convolutional neural networks (CNN) are getting more and more attention in generic object detection [34, 16], and have also attained promising results on people or pedestrian detection [33, 27, 22].

Instead of developing monolithic people detectors, some others made arduous efforts in incorporating contextual information into people detection. Coherent visual patterns, for example, occlusion patterns [15], multi-scale context descriptor [8], and multi-pedestrian patterns [27] have been learned to enhance single human detection. Rather than distinguishing visual modes, furthermore, interactions between human or other objects have also explicitly exploited to leverage spatial or appearance context, and are commonly modeled as structure prediction problem [7, 32, 5]. For instance, Desai *et al.* modeled the spatial co-occurrence of objects in 6 patterns [7], including human hypotheses. Sadeghi *et al.* modeled complex layout of human by introducing visual phrases [32]. Besides, Idrees *et al.* explored locally-consistent scale prior for human detection, using Markov Random Field [18].

3. Our Approach

In this part, we describe a unified framework for detecting people in crowded scenes by leveraging contextual information. Our approach models people interactions by constructing a *context graph*. We first give a mathematical formulation of the problem, and then detail our approach at each step including context graph construction and how to adapt the label propagation technique to address our problem. In Section 4, we will present a greedy-like method to solve the optimization problem described below.

3.1. Problem Formulation

Given an image, let $\mathbb{X} = \{x_i, i = 1 : m\}$ be a set of m human hypotheses generated by an underlying people detector. We purposely set a low detection threshold to allow for more true detections in $\mathbb{X}$, which unfortunately also gets many more undesirable false alarms. Therefore, our task is to find a subset of $\mathbb{X}$ that covers as many as possible true detections, meanwhile with the fewest false alarms brought in. Mathematically, we aim at seeking an indicator vector $\mathbf{Y} = \{y_1, y_2, ..., y_m\}^T \in \{0, 1\}$ (1 means true people detection, otherwise background or other objects) that maximizes a potential function $\Psi(\mathbb{X}, \mathbf{Y})$, such that the visual detections agree on the contextual setting of the image. We define the potential function as follows:

$$\Psi(\mathbb{X}, \mathbf{Y}) = \sum_{i=1}^{m} y_i \psi^u(x_i) + \alpha \sum_{i=1}^{m} y_i \psi^c(x_i, \mathbb{X}, \mathbf{Y}) \quad (1)$$

where $\psi^u(\cdot)$ is the unary potential that utilizes the original detection score of a hypothesis in our case. $\psi^c(\cdot)$ represents the total contextual potentials (support) of a human hypothesis received from others. α is a constant number balancing these two terms. In previous approaches such as proposed in [7, 40, 31], $\psi^c(x_i, \mathbb{X}, \mathbf{Y})$ represents the support of a hypothesis recieved from its neighbors. For example, Desai *et al.* [7] modeled 6 contextual patterns based on relative spatial locations of two hypotheses. While this proves effective in some cases, it lacks a way to model interactions beyond the 6 spatial patterns. Instead of pre-specifying local spatial relationships, in our approach we implicitly model the contextual interaction between any two human hypotheses via a *context graph* $\mathcal{G}$. Thus we have,

$$\psi^c(x_i, \mathbb{X}, \mathbf{Y}) \triangleq \psi^{\mathcal{G}}(x_i, \mathbf{Y}) \quad (2)$$

where $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$ measures how much contextual potential hypothesis x_i can obtain from validated human hypotheses (reflected by $\mathbf{Y}$) based on $\mathcal{G}$. We drop $\mathbb{X}$ in $\psi^{\mathcal{G}}$ here for clarity. While contextual confidence in previous methods [7, 40, 31] can be regarded as a linear combination of interactions between a node and its local neighbors, our graph-based context modeling enables a node to interact with any node in the graph in a nonlinear way through label propagation, effectively and efficiently (see Section 4.1).

3.2. Context Graph

A context graph in our approach is an undirected graph $\mathcal{G} = (V, E)$ used for label propagation, where V corresponds to a set of human hypotheses and E indicates the strength of contextual interaction between any pair of hypotheses. While our focus is to reward human hypotheses contextually consistent to true detections, suppressing false alarms is equally important during label propagation as our input contains a substantial number of errors. For such a purpose, we consider two types of strengthes when constructing $\mathcal{G}$: *attraction* e^+ and *repulsion* e^-. Here *attraction* measures contextual compatibility between two hypotheses while *repulsion* relates to contextual inconsistency.

In our approach, we deliberate 4 types of contextual cues, namely scale, spatial, overlap and social cues, and denote their attraction strengths with e^+_{sc}, e^+_{sp}, e^+_{ov}, and e^+_{so}, respectively. Similarly, the three type of cues involving in repulsion are scale, spatial, and overlap cues, and their repulsion strengths are denoted by e^-_{sc}, e^-_{sp}, and e^-_{ov}.

We further define the overall attraction strength $e^+(i, j)$ between hypothesis x_i and x_j by

$$e^+_{i,j} = \min\left(e^+_{sc}, e^+_{sp}, e^+_{ov}, e^+_{so}\right) \quad (3)$$

and their overall repulsion strength by

$$e^-_{i,j} = \max\left(e^-_{sc}, e^-_{sp}, e^-_{ov}\right) \quad (4)$$

Finally, the context graph is represented by a symmetric matrix $\mathcal{G} \in \mathbb{R}^{m \times m}$, where $\mathcal{G}_{i,j} = e^+_{i,j} - e^-_{i,j}$. To be noticed, $G_{i,j}$ may have negative values, through which repulsion can be spread from true detections to false alarms. The 'min' operation in Eq. (3) implies the attraction of two hypotheses is low whenever one of the contextual attraction is weak. Differently, the'max' operation in Eq. (4) indicates any incompatible pattern should lead to high repulsion.

3.3. Feature Representation of $\mathcal{G}$

We describe below the feature representation for each type of context considered in our paper. These features will be further mapped to a value between 0 and 1 to indicate the strength of the contextual interaction.

Spatial context is explored by the image distance $d(x_i, x_j)$ between two hypotheses. To eliminate the effects of image resolution and camera perspective, $d(x_i, x_j)$ is further normalized as follows: $f_{sp}(x_i, x_j) = \frac{2d(x_i,x_j)}{h_i+h_j} \cdot \max\left(\frac{h_i}{h_j}, \frac{h_j}{h_i}\right)$. Here h_i and h_j are the image heights of hypothesis x_i and x_j. h_i/h_j compensates camera perspective as this ratio can sort of reflect the depth change of the two hypotheses. Here, we use the 'max' operation to obtain symmetric feature. The sum of h_i and h_j further normalizes the distance into the unit of human height.

Scale context is evaluated by the physical height ratio of two hypotheses. We first adopt the method in [17] to estimate the image location of horizon line v_0. By assuming all hypotheses are all grounded and upright, the physical height of hypothesis x_i is defined as $\zeta_i = h_i \zeta_c / (v_i - v_0)$, where h_i and v_i encode the physical height and image location of x_i, respectively. ζ_c is the camera height. Consequently, given any two detections, we can obtain a value pair $(v_0(h_i - h_j), h_i v_j - h_j v_i)$. With multiple ($\geq$ 3)

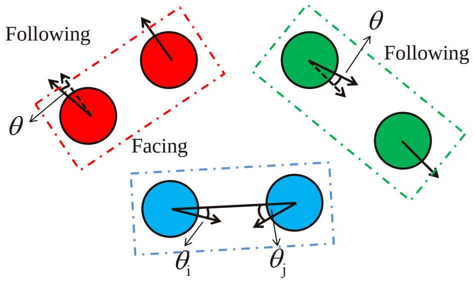


Figure 2. Social context. Each circle represents one human hypothesis; arrows illustrate body (head) orientations. We model two kinds of social interaction patterns between any two hypotheses, i.e., following and facing, which can be indicated by their body (head) orientations.

strong detections (i.e high confidence), v_0 can be easily estimated by least square fitting. The physical height ratio of two hypotheses can then be defined as: $f_{sc}(x_i, x_j) = \min\left(\frac{h_i(v_j - v_0)}{h_j(v_i - v_0)}, \frac{h_j(v_i - v_0)}{h_i(v_j - v_0)}\right)$.

Social context reflects social interactions between two detections measured in virtue of pose and body (or head) orientations. We train a pose classifier (standing vs. sitting) as RBF-kernel SVM with 1200-dim Poselet activation vector (PAV) [23]. The 2-dim probability output p of this classifier is then used to evaluate pose similarity. Body orientations indicate the position of co-existing true detections. We model two orientation patterns in our paper, namely *following* and *facing*. As shown in Fig. 2, considering a pair of hypotheses, $\angle\theta$ represents the included angle of body orientations; $\angle\theta_i$ and $\angle\theta_j$ are the included angles of head orientations and the connecting line of two hypotheses. A small $\angle\theta$ indicates a following pattern, while small $\angle\theta_i$ and $\angle\theta_j$ exhibit a facing pattern. The strongest orientation pattern is used, thus we have $f_{so} = \min(\angle\theta, \max(\angle\theta_i, \angle\theta_j)) \times ||p_i - p_j||_2$. Estimation of body (or head) orientations also takes advantage of PAV [23], which is capable of handling both profile and back views.

Overlap context As conventional approaches [31, 30], we use the overlap ratio of two bounding boxes to express location compatibility of two hypotheses, i.e., $f_{ov} = (B_i \cap B_j)/(B_i \cup B_j)$, where B_i and B_j are bounding boxes of x_i and x_j, respectively.

Model Parameter Fitting We adopt a data-driven approach to learn a mapping function $\mathcal{F} : f \rightarrow e$ for each contextual pattern, using 120 images from Structured Group Dataset *S-GD* [5] that are independent from the evaluation subset.

Basically, we use Gaussian kernel to model mapping functions $\mathcal{F}$, which has been widely used in constructing affinity graph [45, 44]. Parameters of these mapping functions are estimated by fitting the distributions of the 4 context patterns discussed above. We model spatial attraction e^+_{sp} as a 2-component Gaussian distribution, which corresponds to a maximum influence around 0.3 human height and vanishes after 1.5. We set $e^-_{sp} = 1$ when $f_{sp} < 0.1$. For scale attraction e^+_{sc}, we fit it a Gaussian function in the range of $[0, 0.8]$. To counteract height differences of individuals and mild errors of bounding boxes, we set $e^+_{sc} = 1$ when $f_{sc} \in [0.8, 1]$. We further fit features to e^+_{ov}, e^-_{ov}, and e^+_{so} in the similar way.

Note that scale and overlap contexts are *deductive* patterns, i.e., they are discriminative to tell whether a hypothesis is true or not. For instance, if a hypothesis goes against true detections with regard to scale (small e^+_{sc}), then a strong repulsion should be given (large e^-_{sc}). Hence, we have $e^-_{sc} = 1 - e^+_{sc}$ and $e^-_{ov} = 1 - e^+_{ov}$. On the opposite, spatial and social cues are not deductive, suggesting that we can not infer whether or not a hypothesis is invalid, even if it is remote from true detections or no social interactions are observed.

4. Progressive Potential Propagation

Graph-based label propagation has been widely used in semi-supervised learning [45, 36], to perform transductive inference. In our case, we do not have any ‘labeled’ data, so we propose a greedy-like technique, namely progressive potential propagation, to iteratively label hypotheses as true detections and use them for propagation in next run.

4.1. Potential Propagation

Given a context graph $\mathcal{G}$, the first question arising is how the contextual potential $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$ $(i \in [1 : m])$ in Eq. (2) can be obtained. Suppose that we have a label vector $\mathbf{Y} \in \mathbb{R}^m$ for m hypotheses, we first initialize a potential vector $\mathbf{Z} \in \mathbb{R}^m$ as $\mathbf{Y}$. If the contextual potential of hypothesis x_i is targeted, we set $z_i = 0$ to avoid *self-reinforcement*. Under such a setting, a hypothesis x_i can be seemed as labeled when $y_i \neq 0$ and unlabeled otherwise. As aforementioned, strong true detections is more robust in propagating potential, therefore we need to re-weight $\mathcal{G}$. We apply logistic regression to normalize the unary score $\psi^u(x_j) (j \in [1 : m])$ into $w_j \in (0, 1)$, which can be regarded as the true detection probability of hypothesis x_j; we set $w_j = 1$ if $y_j = 1$, since $y_j = 1$ means validated true detection. Then each column of $\mathcal{G}$ is re-weighted by w_j, i.e., $\mathcal{G}'_{\cdot j} = w_j \mathcal{G}_{\cdot j}$. Matrix $\mathcal{G}'$ is further row-normalized such that $\overline{\mathcal{G}}_{ij} = \mathcal{G}_{ij} / \sum_k |\mathcal{G}'_{ik}|$, which is critical for the convergence of the propagation algorithm. We summarize the potential propagation algorithm in Alg. 1.

In line 6, potential is propagated based on $\overline{\mathcal{G}}$ and the potential vector $\mathbf{Z}$ in previous iteration. In line 7, we reset all negative elements in potential vector $\mathbf{Z}$ back to zero, to cut off the potential defused by false alarms. The reason behind is: even there is a strong exclusion between a false positive and a hypothesis, it would be still hard to determine whether the hypothesis is a true detection or a false alarm.

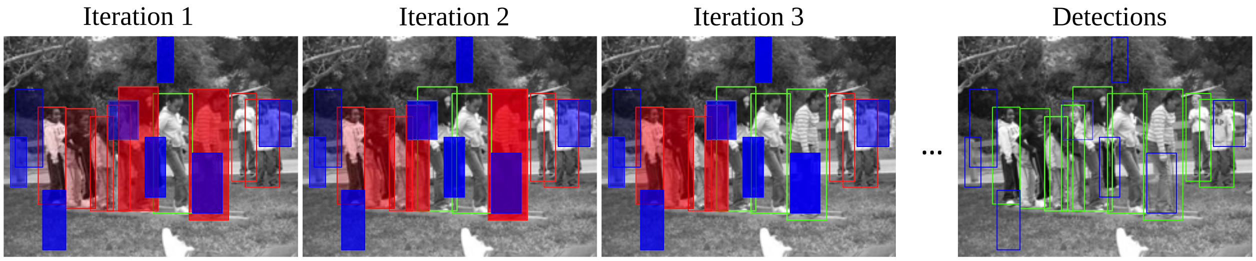


Figure 3. Illustration of progressive inference with potential propagation (image from VOC 2012). Iteration 1 selects the first hypothesis (green bounding box), and propagates contextual potentials (positive or negative) to others in the image. A red bounding box indicates a hypothesis getting positive potentials while a blue one receives negative potentials. Darkness of colors shows the amount of their contextual potentials. After iteration 1, the algorithm picks the hypothesis with the highest potential change composed of both unary and contextual potentials, and then starts the second iteration with 2 instanced human hypotheses. The process repeats until no hypothesis has positive potential gain. In this example, our algorithm ends in 8 iterations, resulting in 8 true detections.

Algorithm 1 Potential propagation for $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$

```
1:  Input: given $\mathcal{G}$,$\mathbf{Y}$.
2:  Output: $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$
3:  Initialize $\mathbf{Z} = \mathbf{Y}$, $z_i = 0$
4:  Obtain $\overline{\mathcal{G}}$ by re-weighting and row-normalization
5:  while $\mathbf{Z}$ does not converge do
6:      $\mathbf{Z} \leftarrow \overline{\mathcal{G}}\mathbf{Z}$
7:      $\mathbf{Z} = \max(\mathbf{Z}, 0)$
8:      $\forall j \in [1:m], j \neq i$, if $y_j = 1$, $z_j = 1$,
9:  end while
10: $\psi^{\mathcal{G}}(x_i, \mathbf{Y}) \leftarrow \overline{\mathcal{G}}\mathbf{Z}$
```

Algorithm 2 Progressive inference for $\hat{\mathbf{Y}}$,$\hat{S}$

```
1:  Input: given $\mathcal{G}$, $\mathbb{X}$
2:  Output: $\hat{\mathbf{Y}}$, $\hat{S}$
3:  Initialize $\mathbf{Y} = 0$ and $S^0 = \emptyset$
4:  First instance $i^* = \arg\max_i \psi^u(x_i)$, set $S^1 = \{i^*\}$.
5:  while $\triangle(x_i) > 0$ do
6:      $\forall i \in [1:m], y_i \neq 1, i^* = \arg\max_i \triangle(x_i)$ (Alg.1)
7:      Update $\mathbf{Y}$: $y_{i^*} = 1$
8:      $S^{t+1} \leftarrow S^t \cup i^*$
9:      $t + 1 \leftarrow t$
10: end while
11: $\hat{\mathbf{Y}} \leftarrow \mathbf{Y}$, $\hat{S} \leftarrow S^t$
```

True detections definitely expose repulsions to false positives. However, since false alarms are heterogeneous, they probably also appear intense exclusive patterns. Similar to [45], in line 8, we replenish elements with initial value 1. The propagation algorithm repeats from line 6 to line 8 until $\mathbf{Z}$ converges. One more propagation is executed in line 10, in order to output the $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$ with both positive and negative contextual potential.

4.2. Progressive Inference

Considering contextual potential $\psi^{\mathcal{G}}(x_i, \mathbf{Y})$, optimizing the objective function in Eq. (1) is NP hard. We thus combine the potential propagation algorithm with greedy forward search, aiming at a sub-optimal $\hat{\mathbf{Y}} = \max_{\mathbf{Y}} \Psi(\mathbb{X}, \mathbf{Y})$. Different from conventional graph-based propagation with fixed 'labeled' instances, we progressively validate unconfirmed hypotheses and propagate their potential in next run. Let $S^t = \{i | y_i = 1, i \in [1:m]\}$ denote the confirmed set of hypotheses at t iteration. We define the potential change by instancing hypothesis x_i as follows:

$$\begin{aligned} &\triangle(x_i) = \Psi(\mathbb{X}, \mathbf{Y}(S^t \cup i)) - \Psi(\mathbb{X}, \mathbf{Y}(S^t)) \\ &= \psi^u(x_i) + \alpha \left(\psi^{\mathcal{G}}(x_i, \mathbf{Y}(S^t)) + \sum_{j \in S^t} \psi^{\mathcal{G}}(x_j, \mathbf{1}(i)) \right) \end{aligned} \quad (5)$$

where $\psi^u(x_i)$ is the unary potential; $\psi^{\mathcal{G}}(x_i, \mathbf{Y}(S^t))$ measures the contextual potentials hypothesis x_i obtains from instanced hypotheses in S^t; $\mathbf{Y}(S^t)$ is the label vector that $y_k = 1$, if $k \in S^t$ and 0 otherwise; $\sum_{j \in S^t} \psi^{\mathcal{G}}(x_j, \mathbf{1}(i))$ represents potentials that hypothesis x_i imposes onto instance(s) in S^t, where $\mathbf{1}(i)$ is an indicator vector with only ith element equals 1 and others are 0s. These two terms can be achieved by using potential propagation algorithm in Section 4.1. Then we devise our progressive inference algorithm as detailed in Alg. 2.

The algorithm starts with an empty set S and a zero vector $\mathbf{Y}$. We select the first hypothesis according to unary potential only. During each iteration, we instance one unconfirmed hypothesis with the largest potential change $\triangle(x_i)$ defined in Eq. (5), and update S^t and $\mathbf{Y}$ accordingly. The algorithm runs line 6 to 9 repetitively and instance one hypothesis in each iteration, until adding any other detections could not enhance the total potential $\Psi(\mathbb{X}, \mathbf{Y})$. Obviously, by growing S^t alternatively, contextual potentials from true detections are progressively propagated. An illustration of the progressive inference with potential propagation is shown in Fig. 3. When the algorithm ceases, the hypotheses in $\widehat{S}$ are regarded as true detections while others as false alarms. We further rescore all detections by summing up their unary and contextual potentials, i.e., we have,

$$\psi'(x_i) = \psi^u(x_i) + \alpha \left(\psi^{\mathcal{G}}(x_i, \hat{\mathbf{Y}}) + \sum_{j \in \hat{S}} \psi^{\mathcal{G}}(x_j, \mathbf{1}(i)) \right) \tag{6}$$

where $\psi'(x_i)$ could be positive or negative. After rescoring, value 0 is further used as the cutoff threshold to differentiate true detections and false alarms (as shown in Fig. 4(a)).

5. Experimental Results

5.1. Datasets and Experimental Setup

Datasets We evaluated our proposed approach on two public datasets: Structured Group Dataset (*SGD*)[1] [5] and ETH pedestrian dataset (*ETH*) [10]. *SGD* shows people with variant poses and layouts while *ETH* contains only pedestrians in low image resolution.

There are a total of 599 images in *SGD*, taken from 6 scenarios (bus stops, classrooms, cafeterias, conferences, libraries, and parks). Crowds in different scenes show various layouts, e.g., queuing, standing in line, sitting in circle, etc. More than $5,000$ people were annotated with tight bounding boxes, torso orientations, and poses (sitting or standing). We randomly chose 20 images from each scenario to form a training subset of 120 images for fitting model parameters (Section 3.3) and learning the pose classifier. Images without sufficient information for horizon line estimation were excluded, leading to 308 images totally for evaluation.

ETH is a standard benchmark for pedestrian detection, containing videos captured in urban settings by a pair of cameras mounted on a chariot. In our experiments, we explored all left video sequences from "Setup 1", which include $1,804$ images and a total of $14,167$ annotated human instances down to a size of about 48 pixels. All these $1,804$ images were used for evaluation.

Experimental Setup Our approach can take input from any detector. On *SGD* dataset, we chose Poselet [4] as the underlying detector for its good ability of handling pose variations. On *ETH*, DPM (LatSvm-V2) [12] was used to verify our proposed method.

Parameter α in Eq. (1) weighs the contextual potential over unary scores. Since detection scores are normally not within the same range across different detectors, α needs to be empirically determined for each underlying detector used in our approach. In our experiments, we set α to 3.0 for Poselet, and 10.0 for DPM, respectively.

For evaluation purpose, we used the popular PASCAL interaction-over-union (IoU) as the measurement to verify the correctness of a human hypothesis. Unless otherwise specified, we set IoU as 0.5 for reporting performance.

5.2. Results

Pose classification and Body Orientation Estimation We first briefly evaluated the performance of pose classification and body orientation estimation on the *SGD* dataset. Over 379 test images, the method [23] achieved an overall accuracy of 84.27% (standing: 81.64%; sitting: 86.41%) for pose classification and a mean error of $20.7°$ for body orientation estimation. The results suggest that useful social cues can be extracted for constructing the contextual graph.

Performance on *SGD* We validated our approach with $6,161$ human hypotheses from Poselet that are above a unary detection threshold of 0.5, and compared the results with those of Poselet [4], and DPM (LatSvm-V2) [12].As shown in Fig. 4(a), our approach outperforms the baseline detector, demonstrating the effectiveness of context modeling. Poselet is improved from 0.669 to 0.684 with regard to average precision (AP).

To further understand the improvement, we reported in Table 1 the precisions and recalls corresponding to the default operational point of our approach (red dot in Fig. 4(a)). At the same recall, our approach yields a much higher precision (78.3% vs. 70.4%) than the baseline detector, i.e., Poselet, while at the same precision, it improves the recall by 4.1% (66.9% vs. 62.8%).

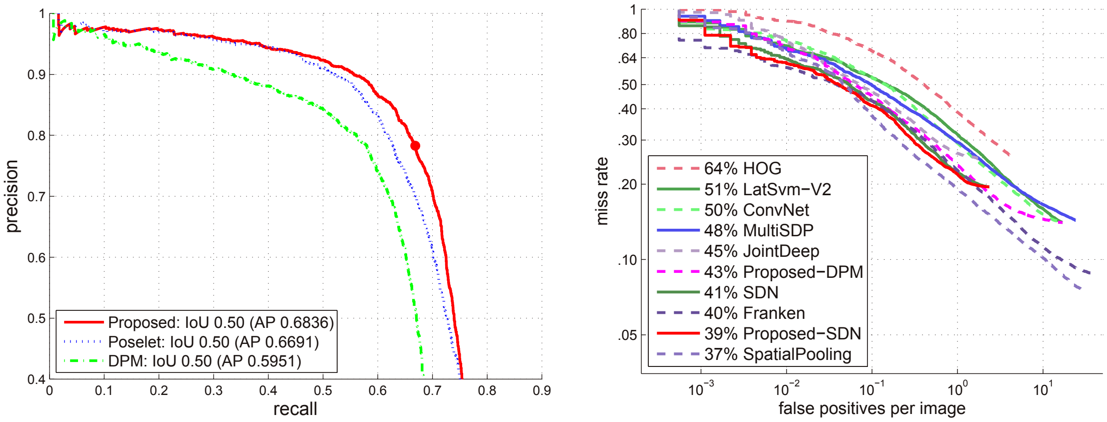
Method	Recall	Precision	F1-score
Proposed	0.6686	0.7829	**0.7212**
Poselet	0.6686	0.7037	0.6857
	0.6280	0.7829	0.6969

Table 1. Evaluation on *SGD*: recalls and precisions of our proposed method at the operational point in comparison with Poselet.

Performance on *ETH* We used DPM as the underlying detector for our approach due to its better performance on *ETH*. The threshold of DPM was set to -0.9, generating a total of $41,744$ pedestrian hypotheses as inputs to our approach. PAV features for predicting pose and body orientations were extracted based on their bounding boxes. In addition to Poselet and DPM (LaSvm-V2), other pedestrian detection methods, such as HOG [6], ConvNet [33], MultiSDP [43], JointDeep [26], Roerei [2], SDN [22], Franken [24], and SpatioPooling [28] were included for comparison, following the evaluation routine in [9][2]. These detectors, expcept HOG and DPM, are top methods proposed recently for pedestrian detection. As can be seen in Fig. 4(b), using DPM as the underlying detector (Proposed–DPM), our approach is as comparative as SDN, achieving a log-average miss rate of 43% and performing better than other deep learning methods including ConvNet, MultiSDP, and JointDeep. The results are encouraging as our method is built

[1] http://cvgl.stanford.edu/projects/groupdiscovery/

[2] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/



(a)

(b)

Figure 4. (a) Recall-precision curves of different approaches on *SGD* dateset; (b) Overall performance on *ETH* dateset in terms of miss rates and false positives per image. The operational point of our approach (at the cutoff threshold of 0.0) is marked as a red dot on the corresponding curve.

Dataset	Baseline	-Scale	-Spatial	-Social	-Overlap	All
SGD	0.6691	0.6632	0.6702	0.6755	0.6620	0.6836

Table 2. Effects of different contextual patterns on average precision. Contextual information was discarded respectively.

upon some well-developed detectors without requiring massive training data and computational training time. We also applied our approach to a strong detector, i.e., Proposed–SDN. The results show that our approach can further reduce the miss rate of SDN from 41% to 39%.

5.3. Discussion

Ablation Study We performed an ablation study on *SGD* dataset to understand the contribution of each contextual cue considered in our approach, in regards of average precision. We singled out one cue each time and reported the performance in Table 2. Apparently, removing any single cue from the context graph leads to deteriorative results, suggesting that all the cues are helpful for people detection. Among them, overlap context contributes most to the final performance as it acts as a primary force in suppressing false alarms.

Benefits of Progressive Inference As described in Section 4.2, our algorithm iteratively picks the best 'true detection' with the largest potential gain at each iteration, and then in the next run uses it as a 'labeled' instance to propagate contextual potential to other unconfirmed hypotheses. To validate the effectiveness of such a progressive fashion, we compared it with a 'threshold-based' method that only does potential propagation once using high-confidence hypotheses, since strong hypotheses are usually associated with true detections.

Method	Proposed	Threshold				
		20	10	5	1	$-\infty$
F1-score	**0.7212**	0.6482	0.6971	0.7115	0.7029	0.6894
Recall	0.6686	0.7293	0.7047	0.6814	0.6036	0.5728
Precision	0.7829	0.5834	0.6896	0.7444	0.8414	0.8656

Table 3. Performance comparisons of our progressive inference and the threshold-based approach on *SGD*, in terms of F1-score at the default operational point.

As can be seen in Table 3, a large threshold (e.g., 20) leads to fewer 'labeled' data samples, thus is incompetent to suppress false alarms (lower precision). Oppositely, a small threshold (e.g., 1) would take many false positives as 'labeled' data that would propagate potentials improperly to true detections (lower recall). As a comparison, our progressive inference is adaptive and can grasp a good tradeoff between precision and recall.

Sampled Detections We illustrated a few sampled detections by our proposed approach in Fig. 5, in which correct detections from the underlying detectors are colored as red while additional detections discovered by our approaches marked as green. These results clearly demonstrates the efficacy of our approach in context modeling. In the second image, we observe a child (marked with a yellow bounding box) incorrectly being suppressed by our approach due to scale inconsistency, since we assume people have roughly equal heights. The last image indicates two false alarms in our approach, which actually highly resemble a human.

6. Conclusions

In this paper we have proposed a novel approach to improve people detection in crowded scenes by exploring

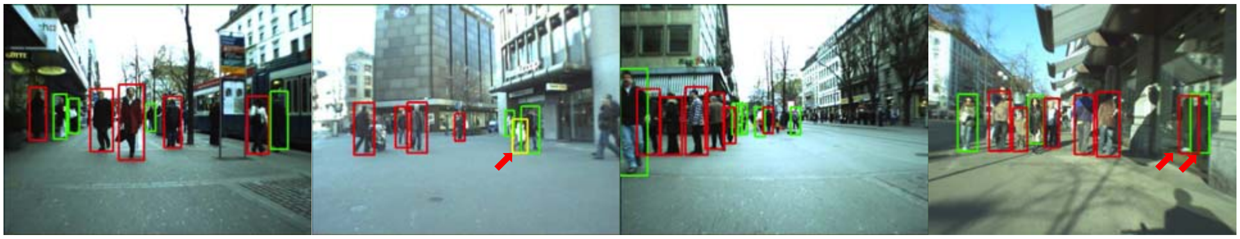


Figure 5. Sampled detections by our proposed approach. Correct detections from the underlying detector are colored as red while detections discovered by our approaches marked as green. Red arrows point out some failed cases in our approach.

contextual cues. Our approach models people interactions through a context graph, via attraction and repulsion built up on both geometric and social cues available in crowded scenarios. Contextual potentials are progressively spread by label propagation, such that contextually compatible human hypotheses get reinforced by receiving positive potentials during the propagation while false alarms are contained due to being negated by contextual incompatibility. We have shown results comparable to state of the art on two public datasets for people and pedestrian detection.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021, 2009.

[2] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In *CVPR*, pages 3666–3673, 2013.

[3] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? *arXiv preprint arXiv:1411.4304*, 2014.

[4] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, pages 168–181, 2010.

[5] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Discovering groups of people in images. In *ECCV*, pages 417–433, 2014.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[7] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 95(1):1–12, 2011.

[8] Y. Ding and J. Xiao. Contextual boost for pedestrian detection. In *CVPR*, pages 2895–2902, 2012.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *T-PAMI*, 34(4):743–761, 2012.

[10] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, pages 1–8, 2007.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *T-PAMI*, 32(9):1627–1645, 2010.

[13] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *CVPR*, pages 1022–1029, 2009.

[14] W. Gao, H. Ai, and S. Lao. Adaptive contour features in oriented granular space for human detection and segmentation. In *CVPR*, pages 1786–1793, 2009.

[15] G. Ghiasi, Y. Yang, D. Ramanan, and C. Fowlkes. Parsing occluded people. In *CVPR*, pages 2401–2408, 2013.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[17] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1):3–15, 2008.

[18] H. Idrees, K. Soomro, and M. Shah. Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *T-PAMI*, 2015.

[19] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011.

[20] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[21] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon. Hierarchical part-template matching for human detection and segmentation. In *ICCV*, pages 1–8, 2007.

[22] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *CVPR*, pages 899–906, 2014.

[23] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, pages 3177–3184, 2011.

[24] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool. Handling occlusions with franken-classifiers. In *ICCV*, pages 1505–1512, 2013.

[25] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou. Discriminative local binary patterns for human detection in personal album. In *CVPR*, pages 1–8, 2008.

[26] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, pages 2056–2063, 2013.

[27] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *CVPR*, pages 3198–3205, 2013.

[28] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *ECCV*, pages 546–561. 2014.

[29] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, pages 452–465. 2010.

[30] R. Rothe, M. Guillaumin, and L. van Gool. Non-maximum suppression for object detection by passing messages between windows. In *ACCV*, 2014.

[31] S. Rujikietgumjorn and R. T. Collins. Optimized pedestrian detection for multiple and occluded people. In *CVPR*, pages 3690–3697, 2013.

[32] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, pages 1745–1752, 2011.

[33] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*, pages 3626–3633, 2013.

[34] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In *NIPS*, pages 2553–2561, 2013.

[35] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *CVPR*, pages 734–741, 2003.

[36] B. Wang, Z. Tu, and J. K. Tsotsos. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *ICCV*, pages 425–432, 2013.

[37] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, pages 32–39, 2009.

[38] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. In *ICCV*, pages 17–24, 2013.

[39] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *CVPR*, pages 90–97, 2005.

[40] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li. Robust multi-resolution pedestrian detection in traffic scenes. In *CVPR*, 2013.

[41] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *T-PAMI*, 35(12):2878–2890, 2013.

[42] C. Yao, X. Bai, W. Liu, and L. J. Latecki. Human detection using learned part alphabet and pose dictionary. In *ECCV*, pages 251–266, 2014.

[43] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In *ICCV*, pages 121–128, 2013.

[44] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *NIPS*, 16(16):321–328, 2004.

[45] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.