Detecting Humans in Dense Crowds Using Locally-Consistent Scale Prior and Global Occlusion Reasoning

Haroon Idrees, *Member, IEEE*, Khurram Soomro, *Member, IEEE*, and Mubarak Shah, *Fellow, IEEE*

Abstract—Human detection in dense crowds is an important problem, as it is a prerequisite to many other visual tasks, such as tracking, counting, action recognition or anomaly detection in behaviors exhibited by individuals in a dense crowd. This problem is challenging due to the large number of individuals, small apparent size, severe occlusions and perspective distortion. However, crowded scenes also offer contextual constraints that can be used to tackle these challenges. In this paper, we explore context for human detection in dense crowds in the form of a locally-consistent scale prior which captures the similarity in scale in local neighborhoods and its smooth variation over the image. Using the scale and confidence of detections obtained from an underlying human detector, we infer scale and confidence priors using Markov Random Field. In an iterative mechanism, the confidences of detection hypotheses are modified to reflect consistency with the inferred priors, and the priors are updated based on the new detections. The final set of detections obtained are then reasoned for occlusion using Binary Integer Programming where overlaps and relations between parts of individuals are encoded as linear constraints. Both human detection and occlusion reasoning in proposed approach are solved with local neighbor-dependent constraints, thereby respecting the inter-dependence between individuals characteristic to dense crowd analysis. In addition, we propose a mechanism to detect different combinations of body parts without requiring annotations for individual combinations. We performed experiments on a new and extremely challenging dataset of dense crowd images showing marked improvement over the underlying human detector.

Index Terms—Crowd analysis, dense crowds, human detection, scale context, spatial priors, locally-consistent scale prior, combinations-of-parts detection, global occlusion reasoning, deformable parts model, Markov Random Field

1 Introduction

Crowd Analysis is fundamental to solving many real-word problems. It is important for management of crowded events, such as protests, demonstrations, marathons, rallies, political speeches and music concerts which are characterized by gatherings of thousands of people. It has use in the design of public spaces and infrastructure, as well as in their expansion and modification, by analyzing the counts of customers and commuters that frequent and travel through these places. It has applications in computer graphics as well, where crowd simulation models can be learned using data from real-world crowded scenes. But, perhaps its most important use is in visual surveillance and anomaly detection. The recurrent and tragic stampedes at pilgrimages [1] and parades [23] as well as the recent terrorist attack at a marathon [5] call for improved and sophisticated techniques for visual analysis of dense crowds.

Since human detection is the primary task even in non-crowded scenes, as it precedes person tracking, action recognition, anomaly detection and higher-level event classification, it has the same importance in crowded scenes where all the other tasks aimed at individuals depend on it. But, a crowd is more than the sum of individuals; the difficulty of computer vision tasks increases disproportionately depending on the number of individuals making up the crowd. This can be gauged by the fact that the human response to an image of a crowd is much slower than that on a non-crowd image. For instance, a human or a member of surveillance team can easily detect, track and count in an image of few people, but when presented with a crowd image containing hundreds to thousands of people, will require a considerably large amount of time. Thus, the straightforward extension of computer vision algorithms does not yield corresponding improvement [41], [49]. And since human detection is fundamental to other tasks in crowd analysis, it assumes even more importance. Until now, the methods that track individuals in dense crowds manually initialize the individuals and track them across frames [3], [28], primarily due to the difficult nature of human detection in crowded scenes.

Dense crowds offer a set of challenges when it comes to visual analysis—fewer pixels per target, perspective effects and severe occlusions. But, they also provide constraints which can be employed to tackle these challenges. These can be both contextual (spatial) or temporal constraints. For instance, tracking methods for dense crowds learn the crowd flow, and use that flow to reliably track individuals in the crowd [3], [28], [37]. Such use of repetition of behavior in time is largely exclusive to dense crowds. In this paper, on the other hand, we explore the use of spatial or

• *The authors are with the Center for Research in Computer Vision (CRCV), University of Central Florida, Orlando, FL 32816. E-mail: {haroon, ksoomro, shah}@eecs.ucf.edu.*

Manuscript received 24 Sept. 2013; revised 4 Jan. 2015; accepted 8 Jan. 2015. Date of publication 22 Jan. 2015; date of current version 4 Sept. 2015. Recommended for acceptance by G. Mori. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2015.2396051

Fig. 1. This figure shows several images from the dataset on which experiments were performed. Although there are severe occlusions and pose variations, the consistency in scale (size of humans) is evident in all images which can be used to restrict the space of detection hypotheses in these images.

contextual constraints for improving human detection. Consider, for instance, the images in Fig. 1 which depict crowds of varying densities. In all the images, it can be observed that the scale or size of neighboring individuals is similar. Furthermore, although the scale changes across all the images, the change in scale is gradual due to the perspective effect and position of the camera which is generally overhead. Even when the camera is not located overhead or there are multiple ground planes, for instance, stairs, stadiums, or multiple head planes, e.g., people sitting and standing, the scale is still locally consistent but with sharp discontinuities due to humans at various heights and depths. These qualitative observations can be embedded in a discontinuity-preserving Markov Random Field that captures the scale of an entire image. Similarly, there can be several heuristic based methods for occlusion handling, but such methods are not applicable in dense crowds, as there are no isolated pairs of individuals which have to be reasoned for occlusion - individual A may occlude B, individual B may occlude C and so on, making all of them tied to each other. An incorrect solution for one individual in a greedy algorithm can affect detection of many other individuals. Thus, instead of developing a greedy or heuristic solution, we leverage the advanced solutions of Integer Programming to simultaneously reason all occlusions and infer visible areas of detections. Unlike, non-crowd human detection methods that detect and localize humans in isolation, our approach solves the problem in a global fashion, thereby, honoring the relationship that individuals in a dense crowd have with each other.

The key ideas presented in this paper are independent of the underlying human detector used, but due to its popularity, we used Deformable Parts Model (DPM) [18] to obtain the scales and confidences of humans and their component parts, which are then used by the proposed method. Since individuals in dense crowds undergo severe occlusions, and full-body human detection is not sufficient to detect all humans, part-based analysis, therefore, assumes greater significance in such scenes. We propose a solution to detect combinations-of-parts of humans which is able to increase recall by detecting partially occluded humans—a common occurrence in dense crowds. This allows us to have multiple detectors that use same parts, which spares us from part-specific annotations and is computationally efficient as it reuses results of filter responses on the shared parts.

Our approach bridges the gap between holistic approaches to crowds and isolated analysis of individuals in non-crowded scenes. The contributions of this paper are summarized as follows: 1) use of locally-consistent scale prior for human detection and an approach for its application in dense crowds, 2) a method to create detectors comprising multiple parts without requiring annotations of those parts, made possible through the use of Latent SVM, 3) occlusion reasoning in crowds with a global solution, and 4) a new and challenging dataset of dense crowd images with tens of thousands of annotated humans.

The rest of the paper is structured as follows. We present literature relevant to our problem in Section 2, followed by technical details of the proposed approach in Section 3. Then, we provide results of our extensive evaluation on the new dataset in Section 4. Finally, we conclude with suggestions for future work in Section 5.

2 Related Work

Human detection is often the precursor to many computer vision tasks and the problem has been tackled by various approaches in literature [8], [12], [13], [16], [18], [26], [29], [35], [43], [47]. A recent comprehensive survey by Dollar et al. [14] compares various state-of-the-art pedestrian detectors and evaluates their performance based on scale, degree of occlusion and localization accuracy. They conclude that under partial occlusion the performance degrades significantly, and becomes *disappointing* at low resolutions. The authors make an assessment that there is still a considerable gap between the current and desired performance of these human detectors. However, they do suggest that the use of some form of context and better occlusion handling can improve the performance of detectors. Another survey from the perspective of traffic safety is by Geronimo et al. [22] which focuses on application of pedestrian detection to assist drivers, with the goal of avoiding possible accidents and casualties.

Human detection poses a range of challenges, the most important of which are to deal with articulation and occlusions. The non-rigid structure and deformity in humans is modeled using the notion of constituent parts which allow certain degree of displacement of parts relative to their desired positions. Several part-based approaches have been proposed in the literature [18], [20], [30], [44], [46]. In [18], the part filters are learned and applied individually, with each part placed relative to the root location and a deformation cost added to the final confidence. Similarly, some of the recent approaches have used the visibility of parts to infer the occluded regions. Ouyang and Wang in a series of papers handle occlusion by modeling visibility of parts as

hidden variables [31], explicit training of multi-person detection [33] or automatically through deep learning [32]. Enzweiler et al. [17] use mixture-of-expert classifiers and train them on features from intensity, depth, and motion to handle partial occlusions. Duan et al. [15] describe the relations between parts using manually defined rules in a hierarchical structure of words, sentences and paragraphs to deal with articulation and occlusions. Wang et al. [43] train a HOG-LBP/SVM classifier, and present a method to find contributions from individual parts, which are used to construct an occlusion map depicting visible regions in detections. Unlike [43] which divides the regular SVM bias among rigid blocks of the human detector for occlusion reasoning, we propose to divide the Latent SVM bias term among deformable parts to automatically create detectors for different combinations of body parts. In particular, we create head, head and shoulder and upper body detectors besides the full body detector using only the annotations for complete humans.

State-of-the-art human detectors perform reasonably well to handle deformation and mild occlusions in non-crowded scenes. However in dense crowds, where individuals undergo severe occlusions, large deformations as well as extreme variations in apparent size, human detection becomes an extremely challenging task. For detection in low-density crowd scenes, a unified probabilistic framework by Yan et al. [48] uses appearance and spatial interaction to describe multiple pedestrians. Improved occlusion handling using a multi-view geometry approach is presented by Ge and Collins [21], who estimate the number of people in a crowd and their locations by sampling from a posterior distribution over a 3d crowd configuration. Similarly, Arteta et al. [4] propose a method to correctly detect overlapping objects using segmentations in video sequences. Crowd density is utilized by Rodriguez et al. [38] who show improved person localization and tracking performance in crowded scenes. They formulate the problem as an optimization of a joint energy function by incorporating confidences of detections subject to overlap and scene-specific density constraints. A video is divided into two sets, where the first set with annotated humans is used to train the density estimator, while the second set is used for testing. The ideas presented in this paper are complementary to [38], but our goal is to use cues or priors that are generally applicable, and not learned from, and applied on, individual scenes or videos.

Human detection is a pre-requisite to tracking, but due to the difficulty in detecting humans in dense crowds, approaches rely on temporal repetition in the form of motion patterns [28], [50] and floor fields [3] to track and analyze crowded scenes, and require manual initialization of tracks. Ali and Shah [3] use this idea in the form of floor fields, which determine the probability of motion from one location to another. Crowd behavior has been similarly modeled by Rodriguez et al. [37] in unstructured scenes to track individuals. Rodriguez et al. [39] use a large collection of crowd videos to learn motion patterns which are then used to drive a tracking algorithm. Multi-target tracking combined with motion pattern learning by Zhao et al. [50] has shown to improve tracking in structured crowds. It requires user labeling of the target in the first frame, which is used to learn a detector, later employed to detect and track other similar objects in the sequence. The common theme in these works is temporal modeling of crowd motion and manual initialization of individual tracks. In addition to tracking, human detection can also be used as a source of information for counting as was done by Idrees et al. [25] who used only the head detections for estimating the number of people in an image. This work differs from [25] as our goal is human detection in terms of bounding boxes and not just counting which produces one number for an entire image. Thus, improved human detection is consequential for counting where it can improve count estimation, as well as for tracking in dense crowds where it can significantly reduce the effort of manual initialization.

Inspired from human visual system which makes use of contextual information to detect and recognize objects, context in computer vision has been extensively studied and used to improve object detection. Researchers have experimented with various approaches: semantics [6], image statistics [42], shape context [36], pixel context [8], [45] and color/texture cues [40], 3D geometric context [24] as well as intensity/depth/motion cues [17]. Divvala et al. [11] evaluate several sources of context and propose the use of geographic context and object spatial support. The work by Desai et al. [9] focuses on learning spatial context to simultaneously predict labeling of a scene while bypassing heuristic-based post-processing steps. Similarly, Ding and Xiao [10] combine the local window with neighborhood windows to construct a multi-scale image context descriptor from HOG-LBP features. Scale information has also been used in different areas, which is either obtained manually or is scene-specific [7], [27], [34] and sometimes requires knowledge of camera parameters [2]. In this paper, we propose to use context in the form of locally-consistent scale prior which enforces the constraint that the size of proximal individuals in a dense crowd is consistent and similar, though there may be occasional discontinuities. Closely related with scale prior is the confidence prior which gives the confidence of the associated scale at each location in the image. We show that both of these priors can be automatically discovered from the scene and are extremely relevant to detecting humans in dense crowds.

3 Framework

In this section, we describe our approach in detail. To keep the paper self-contained, we first describe essential details of Deformable Parts Model [18] that are relevant to our approach, using the same notation as in [18]. We, then, describe how scale and confidence priors are automatically discovered from given images by refining the priors and human detections in an iterative fashion. Next, we present a technique to detect combinations-of-parts using the existing DPM formulation. Finally, the set of putative detections are globally reasoned for occlusion, resulting in bounding boxes on the visible parts of humans as output. Note that the choice of DPM as underlying detector is arbitrary, any human detection algorithm which performs detection at multiple scales and uses part-based models can be substituted in its place.

3.1 Background: Deformable Parts Model

In order to capture the changes in viewpoint as well as variations in pose due to the articulation, Deformable Parts Model [18] uses HOG features to match appearance, and instead of using just the filter scores from rigid templates, it considers deformation, which when represented as a score, measures the displacement of parts from their ideal locations.

To detect objects at multiple scales, a feature pyramid H is constructed with L levels, with $p = (x, y, l)$ representing the position (x, y) at level l in the pyramid. The parameter λ determines the rate for scale sampling in H, i.e., λ is the number of levels down the pyramid at which the resolution doubles compared to a given level. The feature vector at position p in the pyramid is given by $\phi(H, p)$. The appearance score is, then, simply the dot product between filter, F', and feature vector, i.e., $F' \cdot \phi(H, p)$. The model for part i is represented as $P_i = (F_i, v_i, d_i)$, where F_i is the filter for the ith part, v_i is the anchor position w.r.t root position, and d_i is the deformation cost. The deformation score of a part with displacement (d_x, d_y) is given as $d_i \cdot \phi_d(d_x, d_y)$, where ϕ_d returns the deformation features. Finally, an object model with n parts is given by $(F_0, P_1, P_2, \ldots P_n, b)$, where F_0 is the root filter, P_i is the model for ith part consisting of appearance and deformation costs, and b is the constant bias term. The confidence output by human detector, $\mathrm{conf}_{\mathrm{HD}}$, for each hypothesis is the sum of scores from the root filter, filter and deformation scores from the parts, plus the bias, i.e.,

$$\begin{aligned}&\mathrm{conf}_{\mathrm{HD}}(p_0, p_1, p_2, \ldots p_n)\\ &\quad= \sum_{i=0}^{n} F_i' \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(d_{x_i}, d_{y_i}) + b.\end{aligned} \tag{1}$$

3.2 The Scale and Confidence Priors

In a densely crowded image or video, human detection becomes difficult primarily due to the smaller target size and severe occlusions. But, the scale of a human in crowded scene provides cue to what the scale should be in the immediate surrounding of the associated detection. We can transfer the knowledge of scale from a point in scene to its surroundings using the scale and confidence of that particular human detection. Fig. 2 illustrates this idea. Given scale and confidence priors, the confidence for detection hypotheses is altered to reflect conformity with the priors. However, since both the priors and detections are dependent on each other, this necessitates an iterative mechanism where the priors are improved using given detections, and detections are improved using updated priors. Next, we present one cycle of this iterative procedure to discover priors and obtaining the detections.

Inferring scale and confidence priors from given detections. For a detection Ω_q, let (x_q, y_q) denote its position in the image, and s_q and c_q represent the scale and confidence, respectively. Then, given a set of input detections, Ω_q, $q = 1, 2, \ldots, Q$, our goal is to infer the scale and confidence at each location $\mathrm{x} = (x, y)$ in the image. All the detections induce a local influence in terms of scale and confidence, which can be captured with an *Influence Function*, induced by every detection. Such a function should be dependent on locations of input detections since scale-consistency is only valid locally in most images depending on camera location, number of ground planes (stairs, stadiums) or number of head planes (people sitting, standing). It should be a function of scales because a detection with a larger scale has its neighbors at a larger distance than smaller detections. Finally, since the scale information of high-confidence detections is more reliable, it should also be dependent on the confidence, c_q. We propose to use the following function:

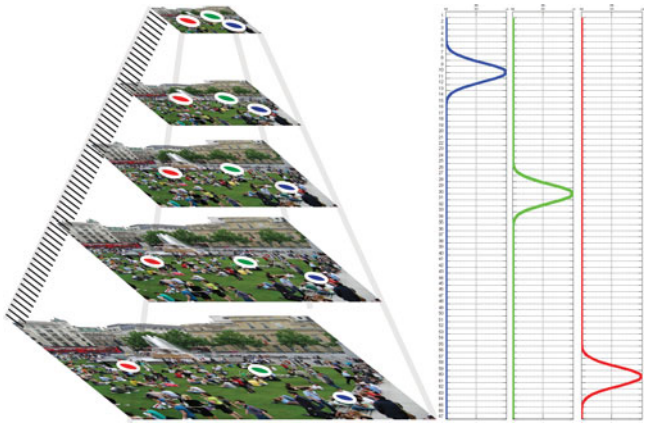


Fig. 2. Human detection in DPM [18] is performed using $L = 67$ levels of the pyramid. Three pixel locations in given image, shown with different colors, have different prior information on scale and confidence. In our approach, the scale and confidence priors are discovered automatically which then provide a 1d scoring function at each pixel in the image, as shown on right. By transforming the priors to each level of the pyramid, the confidence for detection hypotheses is altered based on their consistency with the priors. Increasing the confidence of scale-consistent but low-confidence hypotheses allows them to be detected without incurring false positives in the rest of the image. Effectively, for a 2304×3072 image, this amounts to re-scoring all the 3.85 million hypotheses.

$$\xi_{x,y}(\Omega_q) = c_q \cdot \exp\left(-\frac{\|x - x_q\|^2 + \|y - y_q\|^2}{\sigma^2 \cdot (1 + s_q/\rho(H_{L/2}))^2}\right), \tag{2}$$

where σ is the deviation along x and y axes, and $\rho(H_l)$ returns the scale of a detection at level l in the pyramid. From the above equation, it is evident that $\xi_{x,y}$ is a function of all three aspects of a detection, its location (x_q, y_q), scale s_q and confidence c_q. It also satisfies all the mentioned properties and, therefore, is a valid Influence Function. Furthermore, the detection that has the maximum value at the location (x, y) in the image, Ω_{q*}, determines the value of scale (Θ_s) and confidence (Θ_c) priors at that location, i.e.,

$$\begin{aligned}q* &= \underset{q}{\mathrm{argmax}}\ \xi_{x,y}(\Omega_q), \forall q = 1, 2, 3, \ldots, Q,\\ \Theta_c(x, y) &= \xi_{x,y}(\Omega_{q*}), \quad \Theta_s(x, y) = s_{q*}.\end{aligned} \tag{3}$$

The confidence prior Θ_c at each location in the image is just the maximum value of Influence Function. The scale prior Θ_s is the scale of the particular detection that has the maximum influence at that location. Fig. 3b shows the scale prior for an image shown in Fig. 3a. It is similar to Voronoi Diagram except each region is represented by a scale and the distance-measure is the influence function ξ, instead of the euclidean distance.

Due to perspective effects, the scale of humans changes from pixel to pixel, but its effect is usually gradual. While the humans closer to the camera appear larger, the ones in the background appear smaller. This consistency in scale is

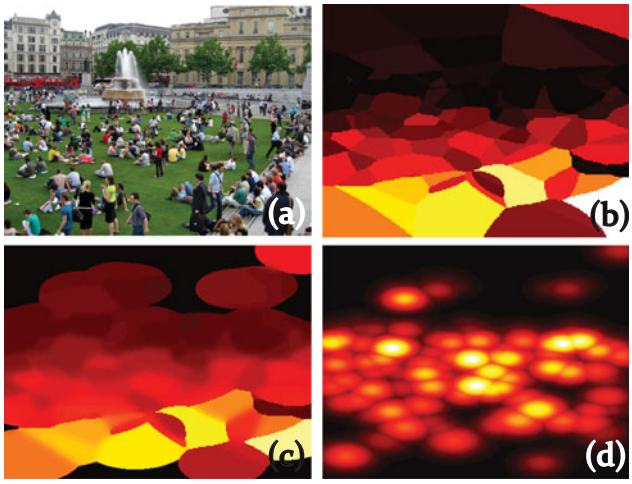


Fig. 3. Intermediate computations of scale and confidence priors: (a) The scales and confidences from detections in an image are transformed into a 2d graph. (b) The observed scale prior is obtained using Equation (3), (c) which is then smoothed through MRF using Equation (4). The corresponding confidence prior is also shown in (d). Heat map is used in (b)-(d) where brighter colors indicate larger values.

imposed by treating scales as random variables and placing them in a Markov Random Field which enforces smoothness at nearby image locations. We model this using grid MRF (inferred using Max-Product/Min-Sum BP [19]), given by,

$$E(\ell) = \sum_{\mathrm{x}\in\mathcal{V}} \Phi_{\mathrm{x}}(\ell_{\mathrm{x}}) + \sum_{(\mathrm{x},\mathrm{x}')\in\mathcal{N}} \Psi(\ell_{\mathrm{x}} - \ell_{\mathrm{x}'}), \tag{4}$$

where Φ, Ψ are the unary and binary potentials and $\mathcal{V}, \mathcal{N}$ define vertices (pixel locations) and neighborhoods in the graph, respectively. The labeling ℓ assigns a label (scale) at every location x in the image. The data term, Φ_{x}, is quadratic, while smoothness term, Ψ, is truncated quadratic. Although the scale varies gradually due to perspective effects, but due to particular viewpoints, there can exist sharp discontinuities. These can also arise from false positives which are likely to be different in scale than correct detections in a particular neighborhood. Thus, it is important to infer the scale prior while preserving the sharp discontinuities. The truncated quadratic cost for smoothness allows us to achieve this objective. Figs. 4a and 4c show the case of rapid scale change and that of scale-inconsistent false positives, respectively. In both cases, the scale information was correctly captured using the proposed approach.

Altering confidences of detection hypotheses given priors. Given priors, the confidences of detection hypotheses are then re-evaluated, as illustrated in Fig. 2. The new confidence is the sum of confidence from the underlying human detector plus the output of scoring function that measures consistency of scale of the detection hypothesis with the scale prior at that location weighed by the confidence prior,

$$\begin{aligned}\mathrm{conf}(\Omega_q) = \mathrm{conf}_{\mathrm{HD}}(\Omega_q) \\ + \alpha \cdot \Theta_c(x_q, y_q) \cdot \exp\left(-\frac{1}{\beta}\|s_q - \Theta_s(x_q, y_q)\|^2\right),\end{aligned} \tag{5}$$

where α, β are the parameters of the scoring function.

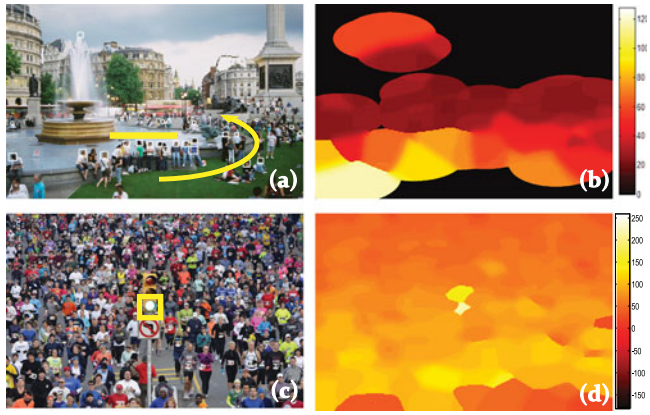


Fig. 4. Intermediate results on inferred scales after smoothing: Two images are shown in (a) and (c) whereas the inferred scale priors are shown in (b) and (d), respectively. Truncated quadratic cost in Equation (4), allows us to handle sharp discontinuities in the scale field, likely to happen at specific viewpoints and due to false positives. The image in (a) has a fountain, where there is a gradual change in scale around it (yellow arrow) but a sharp discontinuity across it (yellow bar). Similarly, in (c), the initial set of detections had a false positive at traffic light larger in size than the immediate neighbors. In both cases, the gradual change in scale and discontinuities were preserved by MRF.

Transformations between priors and feature pyramid. From implementation's perspective, there are two important transformations between scale and confidence priors and each level in the feature pyramid. The first relates the x and y coordinates in priors, (x_Θ, y_Θ), which are the same size as the image, to those in level l in the pyramid, (x_{H_l}, y_{H_l}), and is given by,

$$\begin{bmatrix} x_{H_l} \\ y_{H_l} \end{bmatrix} = \begin{bmatrix} \frac{\rho(H_l)}{\rho(H_{l_0})\cdot k} & 0 & w_0+1 \\ 0 & \frac{\rho(H_l)}{\rho(H_{l_0})\cdot k} & h_0+1 \end{bmatrix} \begin{bmatrix} x_\Theta \\ y_\Theta \\ 1 \end{bmatrix}, \tag{6}$$

where k is the block size used for constructing HOG, w_0, h_0 are the width and height of root filter F_0, and $\rho(H_l)$ is the scale at level l. The second transformation relates the scale in the image or priors to that of each level in the feature pyramid. The $1-1$ mapping that relates size of detection (root template) at image/prior scale to the level l in the pyramid is given by,

$$s = \frac{w_o \cdot h_0 \cdot k}{\rho(H_l)/\rho(H_{l_0})} - 1. \tag{7}$$

In case, we desire to measure the scale of detections in terms of some specific part instead of the root template, for instance the one corresponding to head, we can replace w_0, h_0 with dimensions of that filter in Equation (7), and in image space compute the area of bounding box associated with that part. However, since responses to part filters are computed at twice the resolution, H_l must be replaced with $H_{l+\lambda}$.

The above procedure describes details for one iteration of prior inference and human detection. Fig. 5 quantifies the improvement obtained using the priors at each iteration, as well as the results of baseline [18] and combinations-of-parts detection which is presented in next section. The results are evaluated using only the heads to discard the effect of bounding box sizes. There is very little improvement after third iteration, so we used three iterations in our experiments. The results improve over

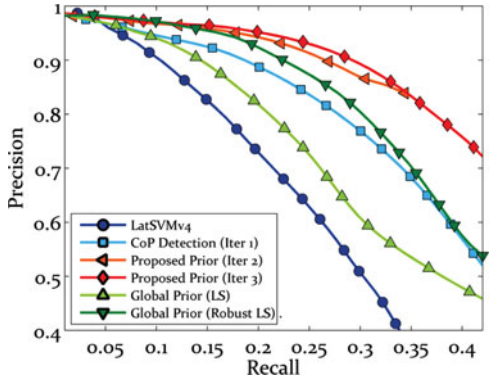


Fig. 5. This graph shows the improvement in performance obtained using the priors over three iterations. The y and x axes show precision and recall, respectively. The curve without the priors is in blue, while the curve with the priors after iterations is in red. Also shown are the results of global scale priors in greens. Details of experimental setup are in Section 4.

Fig. 6. Results after using scale-consistency and combinations-of-parts detection on an image containing almost $3,000$ people. The result is 82 percent precision at 60 percent recall evaluated only on heads. White bounding boxes signify true positives, whereas black represents false positives.

iterations because 1) the correct detections typically have high confidence, and therefore, more influence on their surroundings, 2) the scoring function increases confidence of only those detection hypotheses that are consistent with the scale prior which is smoothed and inferred using many detections from the previous iteration. 3) Although the detections are pre-processed to remove outliers (median filtering) in terms of scale, even if some scale-inconsistent false positive gets through, the discontinuity-preserving MRF ensures that its impact remains restricted. In addition, we report results of global head plane estimation with least squares (LS) in light green and robust LS using RANSAC (RLS) in dark green. Robust LS improves results over baseline but proposed method still outperforms either method of head plane estimation, primarily due to violation of single ground / head plane assumption in many images and re-scoring of all hypotheses before they are selected for output. In Fig. 6, we show how the priors yield impressive results in an image containing extremely dense crowd. For clarity, all detections are shown with the bounding boxes for heads. Interestingly, even the false positives (shown in black) also occur at the correct scale, i.e., at the scale of the neighboring true positives (shown in white).

3.3 Combination-of-Parts (CoP) Detection

Since a human detector always looks for a complete human, it yields low confidences for individuals who are partially visible. To detect such occluded humans in the image, we can lower the threshold, but that incurs false positives which may have higher confidences than the correct but partially visible humans. And since we are dealing with crowded images characterized by severe occlusions, this phenomenon becomes significant. The solution is to detect multiple combinations of parts, which depending on the visibility of parts of an individual, will correspondingly give higher confidence detections. For our approach, we use four different combinations: head $\mathbb{C}_h$, head and shoulders $\mathbb{C}_s$, upper body $\mathbb{C}_u$ and full body $\mathbb{C}_f$. We modified the DPM implementation to detect different combinations of parts by ignoring the filter and deformation scores of excluded parts in each combination. Excluding certain parts affects the Latent SVM bias, since the scores from those parts are not included in the final confidence. In the following treatment, we present a method to divide Latent SVM bias into component parts, which are then used to create CoP detectors.

The bias b in Eq. (1) in Latent SVM [18] is optimized such that confidences of positive examples are greater than 0, while those of negative examples are less than 0. This means that the bias balances the sum of filter and deformation scores from different parts among the positive and negative examples. Therefore, we divide the bias into constituent parts by averaging the contribution of each part to the positive and negative examples while ensuring that the sum of part biases sums to the Latent SVM bias b. Let j and k index positive and negative training examples, respectively. Then, the sum of confidences from the N^+ positive examples using Eq. (1) is given by,

$$\mathcal{S}^+ = \sum_{j=1}^{N^+} \left(\sum_{i=0}^{n} F_i'.\phi(H^j, p_i^j) - \sum_{i=1}^{n} d_i.\phi_d(d_{x_i^j}, d_{y_i^j}) + b \right). \quad (8)$$

Similarly, the sum of confidences from all negative examples is given by,

$$\mathcal{S}^- = \sum_{k=1}^{N^-} \left(\sum_{i=0}^{n} F_i'.\phi(H^k, p_i^k) - \sum_{i=1}^{n} d_i.\phi_d(d_{x_i^k}, d_{y_i^k}) + b \right). \quad (9)$$

Isolating the bias by multiplying Eq. (8) with $\mathcal{S}^-$, Eq. (9) with $\mathcal{S}^+$, and subtracting former from the latter,

$$\begin{aligned}(\mathcal{S}^- N^+ - \mathcal{S}^+ N^-)b \\ = \mathcal{S}^+ \sum_{k=1}^{N^-} \left(\sum_{i=0}^{n} F_i'.\phi(H^k, p_i^k) - \sum_{i=1}^{n} d_i.\phi_d(d_{x_i^k}, d_{y_i^k}) \right) \\ - \mathcal{S}^- \sum_{j=1}^{N^+} \left(\sum_{i=0}^{n} F_i'.\phi(H^j, p_i^j) - \sum_{i=1}^{n} d_i.\phi_d(d_{x_i^j}, d_{y_i^j}) \right).\end{aligned} \quad (10)$$

Now, we simply decompose the bias into parts b_i under the assumption $b = \sum_{i=0}^{n} b_i$. Define $\varrho = \mathcal{S}^- N^+ - \mathcal{S}^+ N^-$. For deformable parts $i \in 1, \ldots, n$ we have,

$$b_i = \frac{\mathcal{S}^+}{\varrho}\sum_{k=1}^{N^-}\left(F_i'.\phi\left(H^k, p_i^k\right) - d_i.\phi_d\left(d_{x_i^k}, d_{y_i^k}\right)\right) - \frac{\mathcal{S}^-}{\varrho}\sum_{j=1}^{N^+}\left(F_i'.\phi(H^j, p_i^j) - d_i.\phi_d(d_{x_i^j}, d_{y_i^j})\right), \tag{11}$$

and for root filter $i = 0$,

$$b_i = \frac{\mathcal{S}^+}{\varrho}\sum_{k=1}^{N^-}\left(F_i'.\phi\left(H^k, p_i^k\right)\right) - \frac{\mathcal{S}^-}{\varrho}\sum_{j=1}^{N^+}\left(F_i'.\phi\left(H^j, p_i^j\right)\right). \tag{12}$$

The bias for CoP detector $\mathbb{C}$ is the sum of bias of its constituent parts, given by $b_{\mathbb{C}} = \sum_{\{i|p_i\in\mathbb{C}\}} b_i$.

The above procedure allows us to detect combinations of different body parts without requiring annotations for them. This advantageous outcome is due to Latent SVM as it infers the location of body parts using training examples when only provided with full-body annotations. Furthermore, although we have found the equivalence between different combinations at zero threshold, $\delta = 0$, the CoP detectors have different sensitivities to changes in δ. For that, we find the linear relationship between confidences of CoP detectors and the full-body detector using the confidences obtained on N^+ positive examples.

3.4 Global Occlusion Reasoning (GOR)

Using the approach described in previous section, we obtain a dense set of CoP detections along with the scores of constituent parts. Although CoP detection significantly improves recall especially for partially visible humans, it adds a layer of complexity to the detection task. On one hand, we can have multiple detections at each location in the image due to outputs from the different CoP detectors. Also since the bounding boxes are placed by each CoP detector without taking into consideration nearby individuals, the resulting detections have significant overlap. On the other hand, it is possible that the bounding box does not cover an individual entirely, due to a relatively higher confidence generated by a CoP detector with fewer parts. Thus, we propose to infer the correct bounding boxes for all the individuals in the scene through occlusion reasoning whose goal is to expand and contract the bounding boxes so that they only but entirely cover the visible parts of the respective individuals. And due to cyclic dependencies among humans in crowds (A occluding B, B occluding C, ...), we pose occlusion reasoning as part-visibility inference problem for all individuals in an image which can be solved in a global fashion through Binary Integer Programming (BIP) given by,

$$\underset{z}{\text{argmin}}\ f^T z \text{ s.t. } Az \le b,\ z \in \{0,1\}^{Qn}, \tag{13}$$

where f contains preferences (based on scores) associated with selecting the corresponding variables in z, which for our problem index over Qn parts from all the Q detections in an image. Independent of the output of CoP detectors, all parts go into the minimization. The scores of parts contributing to CoP detection are increased by the mean confidence over different CoP detections, while rest of the parts are given the raw scores obtained using the full-body detector. Taking the negative of these values gives us the desired f.

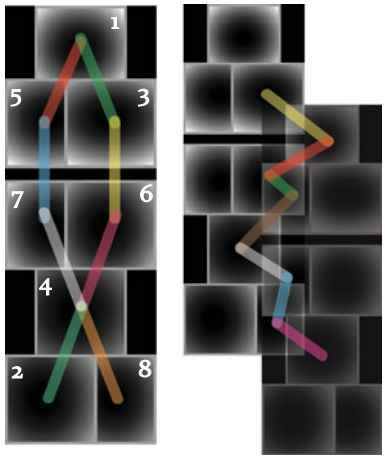


Fig. 7. Linear constraints for Binary Integer Programming: Left shows the DPM model for a single person and the respective part numbers (from model trained on INRIA person dataset). To ensure all parts selected by IP are contiguous, we use chain constraints between parts as shown with different colors. Similarly, models for two occluding persons are shown on right. The overlap constraints ensure that the occluded parts are rejected by the algorithm, thus giving bounding boxes consisting of visible parts only.

However, this would output entire humans whether they are occluded or not. To get non-trivial solutions, we introduce several linear constraints. The first one is based on overlap i.e., if two parts from different individuals have significant overlap, then only one of them should be selected. A single constraint is of the form,

$$\begin{bmatrix}0 & \dots & \mathbf{1}_{o(i,j)>\omega} & \dots & 0 & \mathbf{1}_{o(i,j)>\omega} & \dots & 0\end{bmatrix} z \le \begin{bmatrix}1\end{bmatrix}, \tag{14}$$

where $o(i,j)$ is the overlap of part i with part j from two different detections, obtained by dividing the overlap between i and j with the total area of i and j. The indicator function outputs 1 if overlap is greater than ω. The constraint states that the overlap between two parts which are selected should be less than w, and if it exceeds w, one of the parts should be set to invisible. Each of these constraints forms a row in the matrix A and vector b in Eq. (13).

Overlap constraints alone may result in degenerate solutions, where parts from the head and legs are selected by the optimization while those belonging to shoulders or abdomen are deselected. To alleviate this, we introduce chain constraints which ensure that only contiguous parts of all individuals are selected by the optimization. For a single detection Ω_q, let its corresponding part visibility variables be given by z_q. Using the part numbers given in Fig. 7, the chain constraints are given by $Bz_q \le \mathbf{0}_{7\times 1}$ where,

$$B = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{15}$$

Thus, we have one such set of constraints per detection where each matrix essentially enforces the condition that the part below in a detection should only be selected, if the

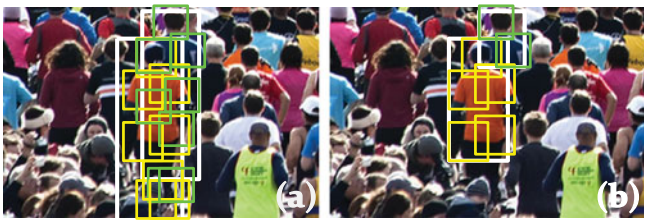


Fig. 8. Results of Occlusion Reasoning: Two individuals are shown in (a) with their bounding boxes for root and deformable parts. (b) After reasoning for occlusion, only visible parts are selected, thus resulting in better localization.

part above is also selected. These constraints are generated automatically by traversing the human model from top to bottom. The relationship between parts 4, 6 and 7 is different from the rest, that is the part 4 can be selected if either part 6 or 7 is selected. For all the detections, chain constraints are written as $\mathsf{diag}(B)z \leq \mathbf{0}_{Qn\times 1}$ where diag operator constructs a block-diagonal matrix using the argument with all other entries set to zero. Finally, the last set of constraints ensure that the output of CoP detectors are not violated. We allow inclusion of new parts immediately below those selected by a CoP detector, for instance, head and shoulders detection $\mathbb{C}_s$ is allowed to become upper-body $\mathbb{C}_u$, whereas the parts further below are hardwired to zero. Although the problem is NP-hard, exact inference is possible for our problem size, as z has $Qn \times 1$ dimensions, and overlap constraints only occur between neighboring detections. We used IBM CPLEX to solve the BIP problem. Fig. 8 shows the results where initial bounding boxes of human and parts are shown in Fig. 8a while results of occlusion reasoning are shown in Fig. 8b.

4 Experiments

We performed experiments on a challenging set of 108 crowd images, downloaded from Flickr. The images cover a variety of scenes and crowd densities, as some are sparse while other are dense. Some of the images depict marathons containing humans in standing poses, while other images are of parks and exhibit more difficult poses. Similarly, severity of occlusion also varies, as in some images, full body detection is possible, while in others, only heads are visible. We manually annotated the images for both heads and visible parts of humans. In total, there are $\sim 35,000$ bounding boxes for head and human each, making UCF-HDDC[1] one of the largest and challenging dataset for Human Detection in Dense Crowds. There are two reasons for annotating heads separately from humans: 1) head bounding boxes can be converted to dotted annotations which mark presence of a human, and thus makes the annotations useful for counting in dense crowds. 2) In dense crowds, heads offer a much better estimate of detection accuracy as evaluation on human bounding boxes quantifies the quality of bounding boxes produced as well. The actual number of human annotations for each image are shown in Fig. 9. This dataset differs from the counting dataset in [25] which only has dotted annotations instead of bounding boxes since the goal in [25] was counting and not detection.

1. http://crcv.ucf.edu/data/UCF_HDDC.php

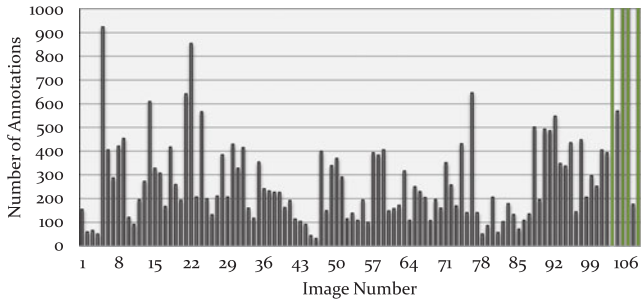


Fig. 9. Statistics on the proposed UCF-HDDC dataset: On *x*-axis is the image id, while *y*-axis shows the number of human annotations in the image. The four green bars at the end have counts of $1,276$, $1,852$, $2,816$ and $2,845$, respectively.

We trained the human detector on INRIA person dataset [8], and used it directly on the proposed dataset. The biases $b_{\mathbb{C}}$ for CoP detectors were also computed on INRIA. For experiments, we used 100 images for testing and eight for validation. The parameters $\alpha = 0.4$, $\beta = 225$ and $\sigma = 300$ were set on the validation set while overlap ratio was arbitrarily defined to $\omega = 0.1$. We used the same value of ω for quantitative evaluation as well, i.e., using 10 percent overlap. The method is robust to changes in MRF and Influence Function parameters, with 50 percent change resulting in 1 percent drop in precision at 40 percent recall. However, 25 percent change in α and β results in almost 4 percent drop in average precision.

4.1 Qualitative Results

Before we present quantitative results, we visualize the improvements obtained by the three ideas presented in this paper in Fig. 10. In this figure, green bounding boxes represent false negatives, black boxes show false positives, while the colored (red to yellow) bounding boxes represent true positives, where brighter colors signify greater overlap with ground truth annotations. The first row shows the gain in performance obtained using the proposed CoP detectors, shown in Fig. 10b, over the full-body human detector, shown in Fig. 10a. Results in both images are shown at 80 percent precision, thus, a higher recall means better performance. The additional humans that were detected are highlighted with yellow arrows. The second row shows the improvement by using scale and confidence priors in addition to CoP detection. The bounding boxes corresponding to heads are shown for clearer visualization. These images also depict results at 80 percent precision, new detections being highlighted by yellow arrows. The third row presents some results of improvement using Global Occlusion Reasoning in addition to CoP detectors and priors. The bounding boxes in the Fig. 10f have much less overlap with each other than those in Fig. 10e. Again, the yellow arrows highlight the improvements—the locations where the occlusion reasoning improved the quality of bounding boxes, the yellow-in-red arrows show the false positives which were removed as a result of reasoning, while red arrows shows a failure case where bounding box became worse. Still, the improvements outnumber the deteriorations, and thus leading to an overall increase in accuracy as suggested by quantitative analysis presented in the next section. Final results on three complete images are shown in Fig. 13.

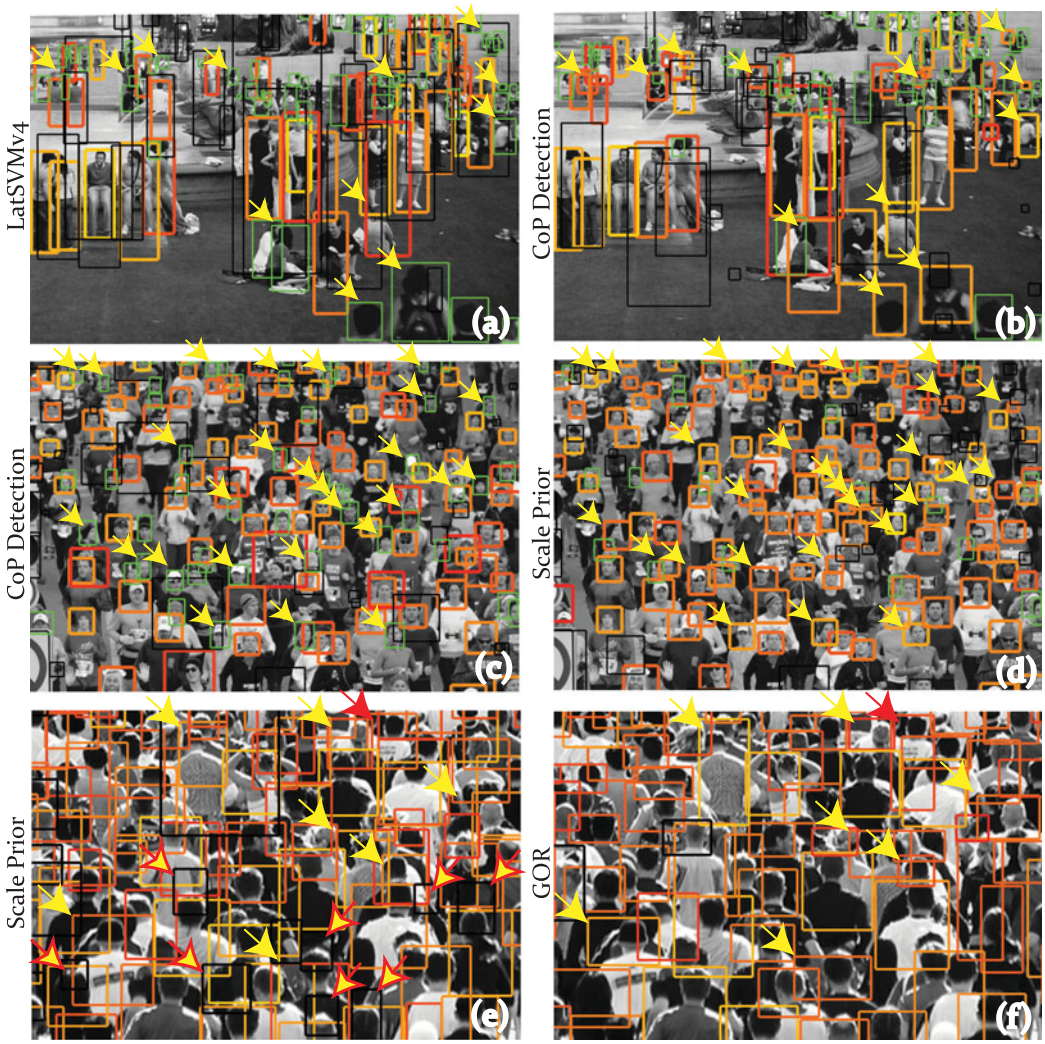


Fig. 10. This figure visualizes the improvements using the three aspects of the proposed approach. Green boxes show false alarms, black represent false positives, while colors in the red to yellow range represent correct detections. In the first row, we show the improvement obtained by using combinations-of-parts detection. Results in both images are shown at the same precision, thus, a higher recall means better performance (shown with yellow arrows). The second row shows gain in performance obtained by using priors in addition to CoP detection. For clarity, only bounding boxes for heads are drawn. Similarly, the last row shows the results of Global Occlusion Reasoning. Yellow arrows indicate improved boxes, yellow-in-red arrows highlight false positives that were removed, while red arrow shows a box that worsened after GOR.

4.2 Quantitative Results

Fig. 11 shows quantitative results of the proposed method evaluated with human bounding boxes using Precision vs. Recall, Miss Rate vs. false positives per image (FPPI), and Multiple Object Detection Precision or MODP. For the first two, we used an overlap of 10 percent, whereas MODP has overlap threshold as the x-axis obtained at 35 percent recall. The first two graphs show that on average, each module of proposed approach improves the performance, with scale and confidence priors and CoP detectors being equally crucial to increase in performance. On the other hand, MODP measures the quality of bounding boxes irrespective of false positives and negatives. The improvement from LatSVMv4 to CoP detectors in terms of MODP is obvious as it is able to pick up occluded humans while incurring fewer false positives. The improvement from CoP detectors to priors is due to change in proportion of true positives to false positives. Since priors reduce the hypotheses space, it reduces the relative number of false positives, which typically occur at random scales and may overlap with annotations. Similarly, since we used exactly the same bounding boxes for occlusion reasoning as were made available after priors, improved MODP suggests that occlusion reasoning results in better localization of detections.

4.3 Density-Based Analysis

To test the robustness and contributions of the three aspects of our method with respect to size of crowd, we performed a density-based analysis in Fig. 12. Here, we simplify the notion of density which refers to the number of humans per image rather than number of people per unit area in real world which is difficult to ascertain. Thus, we sorted the images according to number of annotations, and divided them into four groups: low, medium, high and extreme. In Fig. 12, the first row shows representative images with median counts for each density group. The number of images and some statistics on the number of humans in each group is presented below the median images. Finally, the precision-recall curves are

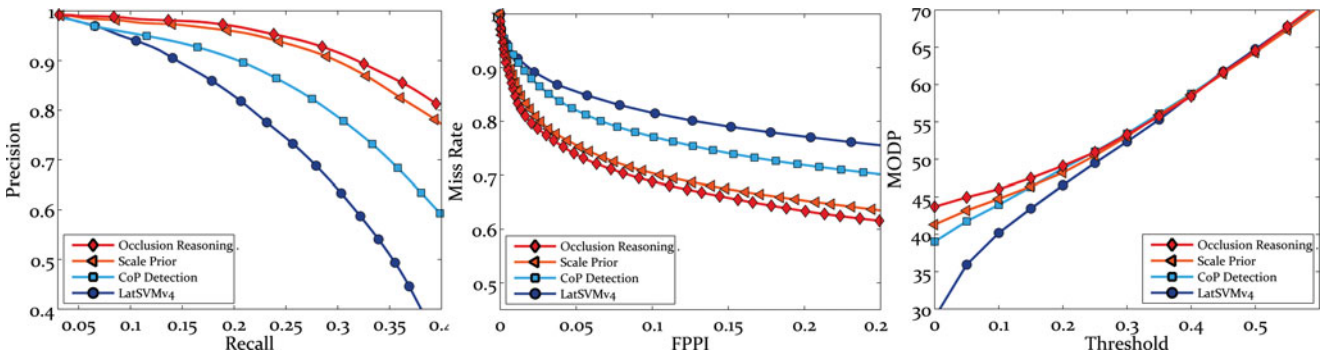


Fig. 11. These graphs show the quantitative results highlighting the contributions of the three different aspects of the proposed approach. Curves in violet-blue show the results of LatSVMv4 [18] (baseline), blue represents CoP detection, orange depicts improvements from priors, whereas red highlights the improvements from Global Occlusion Reasoning.

shown at the bottom of the figure. The curves offer important observations with respect to the three modules. The performance of CoP detection improves with increasing density, simply because humans in high-density undergo more occlusions. The scale and confidence priors give consistent improvement upon CoP detectors across all densities, which is around 15 percent. This means that scale and context is important at all densities. However, occlusion reasoning does not improve at extreme densities, which may be due to the bias of CoP detectors towards combinations with fewer parts for this density. Occlusion reasoning for this group only results in tighter boxes, not affecting the overlap with the predominantly small boxes in annotation, and thus, is not likely to show a noticeable improvement in precision.

4.4 Comparison

We compared the output of the proposed method to several other human detectors. We used the available pre-trained codes provided by the authors. Many work reasonably well in low to high density, but their performance deteriorates on extremely dense images due to severe occlusion. The comparison is shown in Fig. 14 which also shows the Mean Average Precisions along with abbreviated titles.

We used LatSVMv4 [18] as our underlying detection module trained on INRIA person dataset. There are several methods which outperform [18] on this dataset, but still, the proposed approach is able to perform better than all of the other methods. From Fig. 14, we see that at 35 percent recall, the difference between the precisions of proposed and state-of-the-art methods is almost 15 percent. We believe using CN-HOG [26], which is also based on LatSVM, as underlying detection module will further improve the performance of our approach.

Furthermore, it is important to realize that the recall of proposed method is upper-bounded by that of CoP detection, which in turn is dependent on the underlying human detector (LatSVMv4 in our case). It is simply not possible to obtain more detections through the priors or occlusion reasoning than the underlying detection mechanism employed. For this dataset, recall curve hits the asymptote at around 50 percent which is low. Although this is due to the challenging nature of this dataset, we believe in order to obtain better

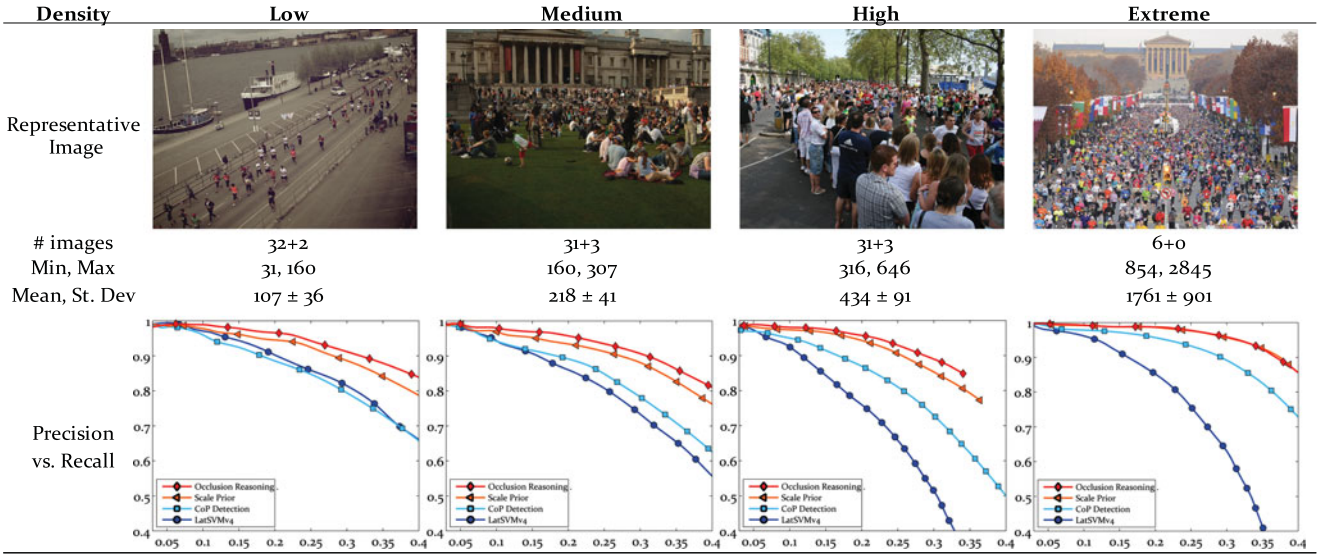


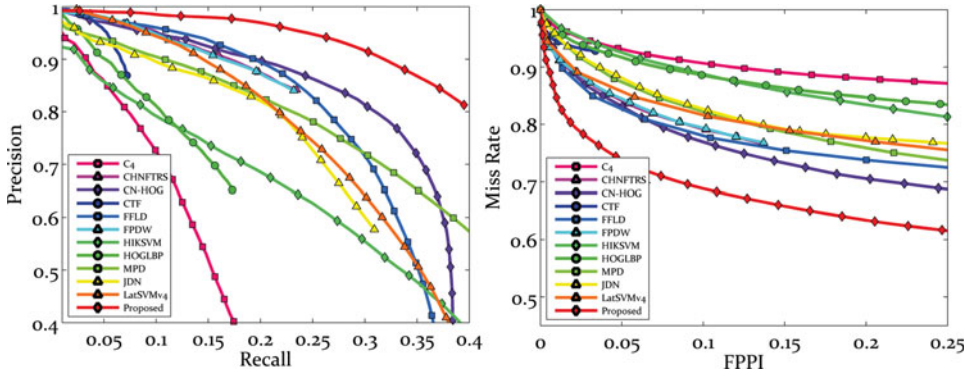
Fig. 12. Density-based analysis: The evaluation on four different densities - low, medium, high and extreme. This figure shows the median image from each density, some statistics on the number of humans, followed by precision-recall curves for LatSVMv4 (baseline), CoP detection, scale and confidence priors and global occlusion reasoning. The addition in # images differentiates images in test and validation set.

Fig. 13. In this figure, white bounding boxes signify true detections (TP), black boxes indicate false alarms (FP), while green represents miss-detections (FN). In (a), the crowd is sparse with humans inclined at an angle due to camera position. In (b), the humans appear in varied poses, whereas (c) is characterized by severe occlusions. The proposed approach gives excellent results for all three scenarios.

performance for human detection in dense crowds, future research must be directed at improving CoP detectors.

4.5 Failure Cases

Due to crowded and challenging nature of the dataset, there are several failure cases. First, we highlight two cases where human or CoP detection is difficult. The first is related to low resolution. Fig. 15a shows a small patch which is $1/400$th the size of original image. Humans in this region become extremely blurred resulting in weak edges and deteriorated HOG-based detection, however the patches are still annotatable by humans. The human size also becomes small, causing issues for DPM as it has a lower limit on detectable part size at 23×23 pixels. The camera position



Method	AP
C4	13.08
CHNFTRS	21.86
CN-HOG	33.62
CTF	6.92
FFLD	31.47
FPDW	21.6
HIKSVM	26.69
HOGLBP	14.42
MPD	32.1
JDN	25.71
LatSVMv4	31.04
Proposed	**37.2**

Fig. 14. This figure shows the comparison of proposed method (red) with several other human detection methods. The methods (training dataset) include C4 (INRIA) [47], CHNFTRS (INRIA) [13], CN-HOG (PASCAL VOC 07,09 & Cartoon) [26], CTF (INRIA, PASCAL VOC 07) [35], FFLD (PASCAL VOC 07) [16], FPDW (INRIA) [12], HIKSVM (INRIA, Daimler-Chrysler) [29], HOGLBP (INRIA) [43], MPD (Extended INRIA) [33], JDN (Caltech, ETH) [32] in addition to LatSVMv4 (INRIA) [18]. The proposed method (INRIA) outperforms all methods on both measures despite using an underlying detector [18] with lower performance than comparison methods.

Fig. 15. Failure Cases: (a) Blurring due to large distance from the camera in addition to small size due to perspective effects. (b) Camera position relative to humans in the scene may result in large number of occlusions where even the heads are partially occluded. (c) Hypersensitivity introduced due to priors caused by large number of high confidence detections in a low density region. (d) High-confidence false positives at the boundary of crowd may results in incorrect initialization of priors. In (c) and (d), true positives are shown with white, while false positives are shown in black.

relative to human height is also important, for instance, in Fig. 15b, even the heads are partially occluded resulting in poor initial detections by the CoP detectors. The solution is to have detectors that are robust to even partial occlusions of parts. Figs. 15c, and 15d show failure cases specific to scale and confidence priors. When we have high-confidence detections in first iteration of prior discovery in a region that has fewer humans per unit area, it sometimes makes the method hypersensitive to detection hypotheses occurring at the desired scale in neighboring areas. This is shown with red arrows in Fig. 15c. Similarly, high confidence non-human detections at early iterations also degrade the scale prior by providing incorrect scale information, thereby resulting in more miss-detections in their surroundings, as can be seen with the balloons in Fig. 15d.

5 Conclusion and Future Work

In this paper, we showed that context, employed in the form of locally-consistent scale and the associated confidence priors, is helpful in improving human detection in dense crowds. Furthermore, we presented an Integer Programming formulation to the task of occlusion reasoning which improves localization of detections. And to detect partially visible humans, we proposed combinations-of-parts detection using different configurations of parts of a complete human. We evaluated our approach using a new set of difficult images, and showed that each aspect is important for detecting humans in dense crowds. Although modularity has its own advantages, it would be an interesting direction to combine all three ideas into one simultaneous solution that also bypasses any post-processing. But, perhaps the most important area of improvement is in human detection itself, where any improvement in the underlying human detector will translate to better performance in dense crowds using the proposed approach. Still this is a new and challenging direction that will have far-reaching consequences to all applications of visual crowd analysis especially safety and surveillance.

Acknowledgments

This material was based upon work supported in part by, the U.S. Army Research Laboratory, the U.S. Army Research Office under contract/grant number W911NF-09-1-0255. Haroon Idrees is the corresponding author.

References

[1] A history of hajj tragedies. (2006). *The Guardian* [Online]. Available: http://www.guardian.co.uk/world/2006/jan/13/saudiarabia

[2] I. Ali and M. N. Dailey, "Head plane estimation improves the accuracy of pedestrian tracking in dense crowds," in *Proc. 8th Int. Conf. Control Autom. Robot. Vis.*, 2010, pp. 2054–2059.

[3] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 1–14.

[4] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, "Learning to detect partially overlapping instances," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3230–3237.

[5] T. E. Board. (2013, Apr. 15). Bombs at the marathon. *The New York Times* [Online]. Available: http://www.nytimes.com/2013/04/16/opinion/bombs-at-the-boston-marathon.html?ref=bostonmarathon

[6] P. Carbonetto, N. Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *Proc. 8th Eur. Conf. Comput. Vis.*, 2004, pp. 350–362.

[7] A. B. Chan, Z.-S. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–7.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2005, pp. 886–893.

[9] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," *Int. J. Comput. Vis.*, vol. 95, no. 1, pp. 1–12, 2011.

[10] Y. Ding and J. Xiao, "Contextual boost for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2895–2902.

[11] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1271–1278.

[12] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 1–68.

[13] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–91.

[14] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[15] G. Duan, H. Ai, and S. Lao, "A structural filter approach to human detection," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2010, pp. 238–251.

[16] C. Dubout and F. Fleuret, "Exact acceleration of linear object detectors," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2012, pp. 238–251.

[17] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 990–997.

[18] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[19] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, 2006.

[20] M. Fink and P. Perona, "Mutual boosting for contextual inference," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 1515–1522.

[21] W. Ge and R. Collins, "Crowd detection with a multiview sampler," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 324–337.

[22] D. Geronimo, A. Lopez, A. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.
[23] D. Helbing and P. Mukerji, "Crowd disasters as systemic failures: Analysis of the love parade disaster," *EPJ Data Sci.*, vol. 1, no. 1, pp. 1–40, 2012.
[24] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2137–2144.
[25] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2547–2554.
[26] F. Khan, R. Anwer, J. Weijer, A. Bagdanov, M. Vanrell, and A. Lopez, "Color attributes for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3306–3313.
[27] D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *Proc. 18th IEEE Int. Conf. Pattern Recog.*, 2006, pp. 1187–1190.
[28] L. Kratz and K. Nishino, "Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 987–1002, May. 2012.
[29] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.
[30] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. 8th Eur. Conf. Comput. Vis.*, 2004, pp. 69–81.
[31] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3258–3265.
[32] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2056–2063.
[33] W. Ouyang and X. Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3198–3205.
[34] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 241–254.
[35] M. Pedersoli, A. Vedaldi, and J. Gonzalez, "A coarse-to-fine approach for fast deformable object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1353–1360.
[36] D. Ramanan, "Using segmentation to verify object hypotheses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[37] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in unstructured crowded scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1389–1396.
[38] M. Rodriguez, I. Laptev, J. Sivic, and J. Audibert, "Density-aware person detection and tracking in crowds," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2423–2430.
[39] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert, "Data-driven crowd analysis in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1235–1242.
[40] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human detection using partial least squares analysis," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, 2009, pp. 24–31.
[41] J. Silveira, J. Junior, S. Musse, and C. Jung, "Crowd analysis using computer vision techniques," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 66–77, Sep. 2010.
[42] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.
[43] X. Wang, T. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, 2009, pp. 32–39.
[44] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1705–1712.
[45] L. Wolf and S. Bileschi, "A critical view of context," *Int. J. Comput. Vis.*, vol. 69, no. 2, pp. 251–261, 2006.
[46] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, 2005, pp. 90–97.
[47] J. Wu, C. Geyer, and J. Rehg, "Real-time human detection using contour cues," in *Proc. Int. Conf. Robot. Autom.*, 2011, pp. 860–867.
[48] J. Yan, Z. Lei, D. Yi, and S. Li, "Multi-pedestrian detection in crowded scenes: A global view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3124–3129.
[49] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, and L. Xu, "Crowd analysis: A survey," *Mach. Vis. Appl.*, vol. 19, nos. 5/6, pp. 345–357, 2008.
[50] X. Zhao, D. Gong, and G. Medioni, "Tracking using motion patterns for very crowded scenes," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 315–328.

Haroon Idrees received the BSc (Honors) degree in computer engineering from the Lahore University of Management Sciences, Pakistan, in 2007, and the PhD degree in computer science from the University of Central Florida in 2014. He is a postdoctoral associate at the Center for Research in Computer Vision at the University of Central Florida. He has published several papers in conferences and journals such as CVPR, ECCV, *Journal of Image and Vision Computing*, and *IEEE Transactions on Pattern Analysis and Machine Intelligence*. His research interests include crowd analysis, object detection, visual tracking, multi-camera and airborne surveillance, and multimedia content analysis. He is a member of the IEEE.

Khurram Soomro received the BSc (Honors) and MSc degrees in computer engineering from the Lahore University of Management Sciences, Pakistan, in 2007 and 2011, respectively. He joined the Center for Research in Computer Vision, University of Central Florida in 2011, where he is currently working toward the PhD degree in computer vision. His research interests include action recognition and localization, human detection, visual surveillance and tracking, and sports analytics. He is a member of the IEEE and Upsilon Pi Epsilon (UPE) Honor Society.

Mubarak Shah is the trustee chair professor of computer science and the founding director of the Center for Research in Computer Vision at the University of Central Florida (UCF). He is an editor of an international book series on video computing, editor-in-chief of *Machine Vision and Applications Journal*, and an associate editor of *ACM Computing Surveys Journal*. He was the program cochair of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2008, an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and a guest editor of the special issue of the *International Journal of Computer Vision* on Video Computing. His research interests include video surveillance, visual tracking, human activity recognition, visual analysis of crowded scenes, video registration, UAV video analysis, and so on. He is an ACM distinguished speaker. He was an IEEE distinguished visitor speaker for 1997-2000 and received the IEEE Outstanding Engineering Educator Award in 1997. In 2006, he was awarded a Pegasus Professor Award, the highest award at UCF. He received the Harris Corporation's Engineering Achievement Award in 1999, TOKTEN Awards from UNDP in 1995, 1997, and 2000, Teaching Incentive Program Award in 1995 and 2003, Research Incentive Award in 2003 and 2009, Millionaires Club Awards in 2005 and 2006, University Distinguished Researcher Award in 2007, Honorable mention for the ICCV 2005 Where Am I? Challenge Problem, and was nominated for the Best Paper Award at the ACM Multimedia Conference in 2005. He is a fellow of the IEEE, AAAS, IAPR, and SPIE.