Density-aware person detection and tracking in crowds

Mikel Rodriguez, Ivan Laptev, Josef Sivic, Jean-Yves Audibert

Density-aware person detection and tracking in crowds

Mikel Rodriguez[1,4] Ivan Laptev[2,4] Josef Sivic[2,4] Jean-Yves Audibert[3,4]

[1]École Normale Supérieure [2]INRIA [3]Imagine, LIGM, Universite Paris-Est

Abstract

We address the problem of person detection and tracking in crowded video scenes. While the detection of individual objects has been improved significantly over the recent years, crowd scenes remain particularly challenging for the detection and tracking tasks due to heavy occlusions, high person densities and significant variation in people's appearance. To address these challenges, we propose to leverage information on the global structure of the scene and to resolve all detections jointly. In particular, we explore constraints imposed by the crowd density and formulate person detection as the optimization of a joint energy function combining crowd density estimation and the localization of individual people. We demonstrate how the optimization of such an energy function significantly improves person detection and tracking in crowds. We validate our approach on a challenging video dataset of crowded scenes.

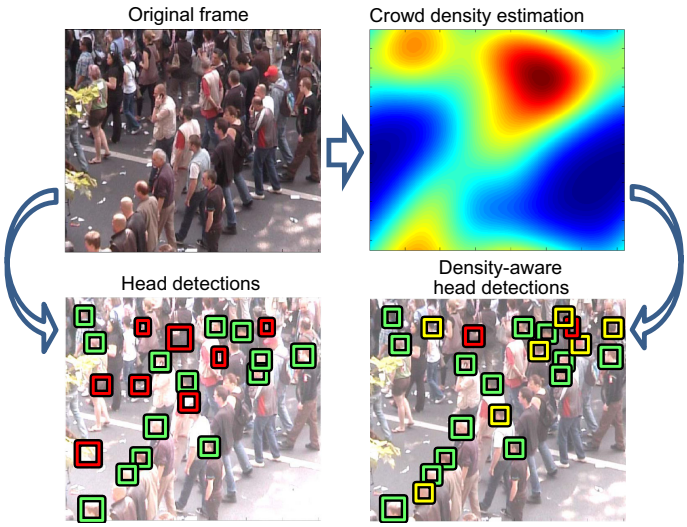


Figure 1. Individual head detections provided by state-of-the-art object detector [12] (bottom-left; green: true positives; red: false positives) are improved significantly by our method (bottom-right; yellow: new true positives) using the crowd density estimate (top-right) obtained from the original frame (top-left).

1. Introduction

Detecting and tracking people in crowded scenes is a crucial component for a wide range of applications including surveillance, group behavior modeling and crowd disaster prevention. The reliable person detection and tracking in crowds, however, is a highly challenging task due to heavy occlusions, view variations and varying density of people as well as the ambiguous appearance of body parts, e.g. the head of one person could be similar to a shoulder of a near-by person. High-density crowds, such as illustrated in Figure 1, present particular challenges due to the difficulty of isolating individual people with standard low-level methods of background subtraction and motion segmentation typically applied in low-density surveillance scenes.

In recent years significant progress has been made in the field of object detection and recognition [7, 11, 12]. While standard "scanning-window" methods attempt to localize objects independently, several recent approaches extend this work and exploit scene context as well as relations among objects for improved object detection [8, 25, 29, 31]. Related ideas have been investigated for human motion analysis where incorporating scene-level and behavioral factors effecting the spatial arrangement and movement of people have been shown beneficial for achieving improved detection and tracking accuracy. Examples of explored cues include: the destination of a pedestrian within the scene [23], repulsion from near-by agents due to the preservation of personal space and social grouping behavior [4], as well as the speed of an agent in the group [15].

We follow this line of work and extend it to the detection and tracking of people in high-density crowds. Rather than modeling individual interactions of people, this work exploits information at a more global level provided by the crowd density and scene geometry. Crowd density estimation has been addressed in a number of recent works which often pose it as a regression problem [5, 17, 18]. Such methods avoid the hard detection task and attempt to infer person counts directly from low-level image measurements, e.g. histograms of feature responses. Such methods, hence, provide person counts in image regions but are uncertain

[4]WILLOW project, Laboratoire d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548.

about the location of people in these regions. This information is complementary to the output of standard person detectors which optimize precise localization of individual people but lack the global knowledge on the crowd structure. Our *goal and contribution* is to combine these two sources of complementary information for improved person detection and tracking. The intuition behind our method is illustrated in Figure 1 where the constraints of person counts in local image regions help improving the standard head detector.

We formulate our method in an energy minimization framework which combines crowd density estimates with the strength of individual person detections. We minimize this energy by jointly optimizing the density and the location of individual people in the crowd. We demonstrate how such optimization leads to significant improvements of state-of-the-art person detection in crowded scenes with varying densities. In addition to crowd density cues, we explore constraints provided by scene geometry and temporal continuity of person tracks in the video and demonstrate further improvements for person tracking in crowds. We validate our approach on challenging crowded scenes from multiple video datasets.

The rest of the paper is organized as follows. We put our method in the context of related work in Section 2. Section 3 describes our model, its individual components as well as the optimization procedure for our objective function. Finally, the proposed model is experimentally evaluated in Section 4.

2. Related work

Object recognition and localization has been studied extensively in computer vision and substantial progress has been achieved over the last decade [7,11,12]. Recent studies demonstrate that the explicit modeling of scene-level information can aid performance of individual object-level detectors. For example, [29] uses scene priming to constrain locations of objects in the image, [3, 30] seek detections that explain a set of detected parts, whereas [8, 25] model spatial and co-occurence interactions between objects in the scene while [31] attempts to explore relations between human poses and manipulated objects. Our detection method relates to this previous work and extends it by exploring the new scene-level constraint provided by the crowd density.

In the area of tracking, [4, 16, 22, 23] demonstrate how the tracking accuracy can be improved by the joint modeling of motion of multiple people in the scene. For example, [22] tracks football players and explores the roles of players to resolve tracking ambiguity. Others study tracking in crowded scenes containing, for example, groups of ants [16] or biological cells [19]. Given the difficulty of tracking individuals in crowds, several approaches leverage common motion patterns as constraints for the tracking problem. In [2] global motion patterns are learned and participants of the crowd are assumed to behave in a similar manner. Overlapping motion patterns have been studied in [26] as a means of coping with multi-modal crowd behaviors and [27] uses a large database of crowds as a data-driven prior for tracking. Our method is complementary to this prior work and can benefit from its sophisticated motion models. Differently to the prior work on high-density crowd tracking, our method builds on the tracking-by-detection framework in [10] and does not require manual initialization.

The problem of determining the density of objects in a scene has been also studied extensively. Several density estimation methods are based on aggregating counts obtained from local object detectors. These methods typically assume that image regions of target objects have distinct motion and/or appearance [9, 24]. While this assumption may be applicable to low- and medium-density crowds, it usually does not hold for high-density scenes. An alternative commonly used density estimation approach is based on regression. These methods [5,6,17,18,20] typically forego the challenges of detecting individual agents and instead focus on learning a mapping between density and a set of feature responses. In our work we apply the regression-style crowd density estimation approach and use it to improve localization of individual people.

3. Crowd Model

This section presents our method for detecting and tracking people in crowds. Section 3.1 formulates the new model for density-informed person detection. The components of this model, i.e. person detection and crowd density estimation are described in Sections 3.2 and 3.3 respectively. Section 3.4 describes our tracking framework.

3.1. Energy formulation

We formulate the density-informed person detection as follows. We assume to have a confidence score $s(p)$ of a person detector for each location $p_i, i = 1...N$ in an image. In addition, we assume we are given a person density, i.e. the number of people per pixel, $D(p_i)$ estimated in a window of size σ at each location p_i. Our goal is to identify locations of people in the image such that the sum of detector confidence scores at those locations is maximized while respecting the density of people given by D and preventing significantly overlapping detections, i.e. detections with the area overlap greater than a certain threshold. Using similar notation as in [8], we encode detections in the entire image by a single N-vector $x \in \{0, 1\}^N$, where $x_i = 1$ if the detection at p_i is "switched on" and 0 otherwise. The detection problem can be then formulated as a minimization of

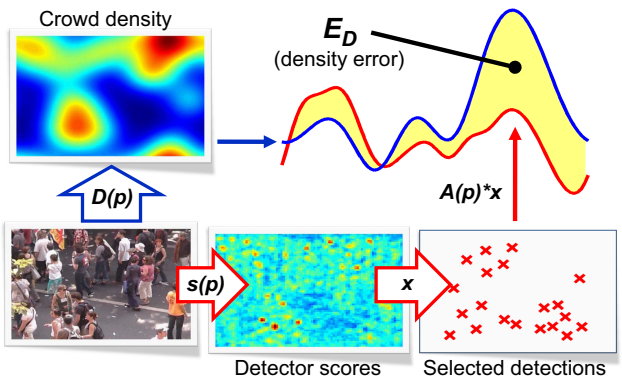


Figure 2. Illustration of the energy term E_D of (1). Minimizing E_D implies reducing the difference in person density estimates obtained by the estimator $D(p)$ (blue color coding) and by local counting person detections (red color coding).

the following cost function

$$\min_{x\in\{0,1\}^N} -\underbrace{s^\top x}_{E_S} + \underbrace{x^\top W x}_{E_P} + \underbrace{\alpha\|D - Ax\|_2^2}_{E_D}. \qquad (1)$$

Minimizing the first term, E_S, in (1) ensures the high confidence values of the person detector at locations of detected people (indicated by $x_i = 1$). The second, pairwise, term E_P ensures that only valid configurations of non-overlapping detections are selected. This is achieved by setting $W_{ij} = \infty$, if detections at locations p_i and p_j have significant area overlap ratio, and 0 otherwise. The first two terms of the cost function are similar in spirit to the formulation used in [8] and implement a variation of the standard non-maximum suppression. In addition, we introduce a new term, E_D, that models the crowd density constraints by penalizing the differences between the density values (i) measured with a regression-based density estimator D and (ii) obtained by counting the "switched on" (or *active*) detections in x. The evaluation of the density of active detections in x is performed by matrix multiplication Ax, where A is a $N \times N$ matrix with rows A_i

$$A_i(q_j) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|p_i - q_j\|^2}{2\sigma^2}\right) \qquad (2)$$

corresponding to Gaussian windows of size σ centered at positions p_i. The idea of minimizing the term E_D is illustrated in Figure 2. Intuitively, optimizing the cost (1) including the third, density, term E_D enables improving person detection by penalizing confident detections in low person density image regions while promoting low-confidence detections in high person density regions.

In the simplest form, when σ is equal to the width of the entire image, the density term E_D reduces (up to a multiplicative constant) to penalizing the difference between the total number of switched on detections in x and the total number of people in the image obtained by the density estimator D.

Note that results of the optimization will be naturally dependent on the quality of the used detector s and density estimator D. Moreover, the quality of D is expected to decrease for small window sizes σ, since if the density estimates D were reliable for arbitrary small σ, having D would solve the detection problem. To balance the contributions of person detection and density estimation, we introduce in (1) a weighting parameter α, which we set manually during training. We also select appropriate window size σ resulting in reasonable density estimates as will be explained in Section 3.3.

Optimization. Optimizing (1) with respect to a discrete vector x is a NP-hard problem in general. As the cost function (1) has a quadratic form, one could attempt optimizing it using quadratic programming by relaxing the binary values of x to $x \in [0,1]^N$. As the quadratic form is not convex (because of the E_P term), there is no real gain in the computational complexity to solve the relaxed problem. We thus follow a different approach and optimize (1) using a greedy search procedure similar to [8]. We initialize $\hat{x} = 0$, then at each new iteration we update $\hat{x}$ with $x_i = 1$ by activating a detection at p_i for i which decreases the cost function by the largest amount. We continue iterations until the cost in (1) cannot be decreased further by activating more detections. Similar to [8], we found this minimization strategy to work well in practice.

3.2. Person detection

In this section we describe the multi-scale person detector, which provides the dense detection score map $s(p)$ used in the cost function (1). Person detection is a widely studied problem with many methods and evaluation benchmarks available [7, 11, 12]. Most of the methods consider full-body (pedestrians) or upper-body detection as these views are typical for streetview scenes, movies and consumer photographs. In dense crowd scenes, however, the body of the person does not provide reliable cues due to frequent occlusions by other people. In this work we use the state-of-the-art object detector [12] and train it on extended head regions obtained by manual annotation of head rectangles in our training images of crowds. To improve discrimination, we sample negatives from the same images by allowing a small overlap with positive samples. Once trained, the detector is applied to all positions and scales in the image to provide a dense score map $s(p)$. To localize individual people a non-maximum suppression procedure is usually applied to $s(p)$ [12]. In our work we use $s(p)$ in combination with density estimation $D(p)$ in (1). Results of person detections in crowds are illustrated in Figures 1 and 4, and are quantitatively evaluated in Section 4.

Geometric consistency. Footage of crowd scenes typically involves people moving on a ground plane filmed with

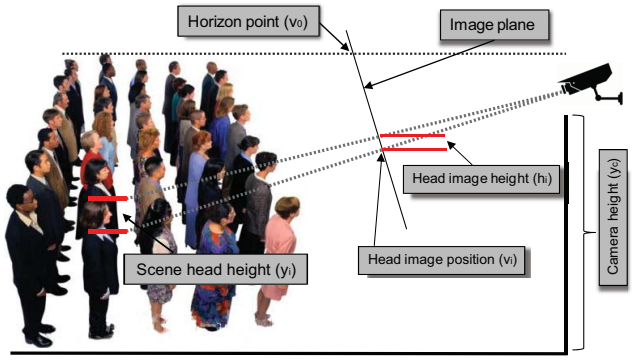


Figure 3. Estimating approximate camera-scene geometry from head detections. Assuming known average head size in the scene, the camera height and tilt (parametrized by the location of the horizon line) can be computed from imaged positions and scales of two head detections in the image. See text for more details.

an overhead camera (see Figure 3). This imposes strong prior on the plausible positions and sizes of people in the frame.

We use this prior to improve results of the basic person detector described above. For this purpose, we seek to roughly estimate the camera viewpoint, which in turn will give us constraints on person sizes observed across the image. We follow the approximation of scene-camera geometry developed in [14] and [13] for generic object detection and adapt it to constrain head detection in crowd videos. We assume people stand on the ground plane, and further assume that (bottom tip of) their heads lie in a single "head plane", parallel with the ground plane. The camera position with respect to the scene is parameterized by two variables: the height of the camera, y_c, and the camera tilt, parameterized by the position of the horizon in the image v_0. We assume there is no horizontal or in-plane rotation of the camera. Given this simplified model, the 3D head size y_i of a person i is related to its measured image height h_i at vertical image co-ordinate v_i as

$$y_i = \frac{h_i y_c}{v_i - v_0}. \tag{3}$$

Assuming known 3D head size, equal to the average person head (22.6cm [21]), we can estimate the two unknown camera parameters (y_c, v_0) from positions and heights of two detections in the image. In our experiments, we estimate the camera parameters using least squares fit from the top m head detection hypothesis with the highest confidences (see figure 4). While the scene-camera geometry model is only approximate, we show that the resulting geometric constraints can significantly improve the performance of the detector in Section 4.

Having obtained an estimate for the camera parameters, we can in turn use them to prune head detection hypotheses which are not consistent with the geometry of the scene. In particular, given the combined detection score map obtained

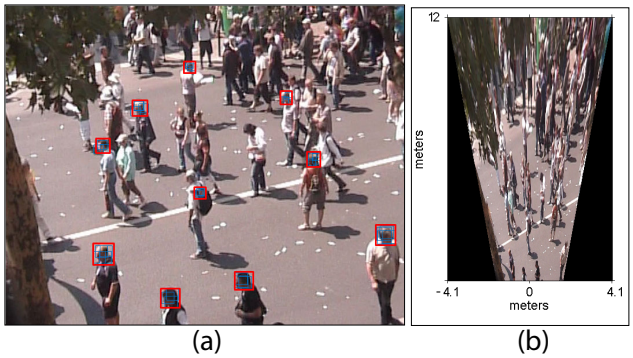


Figure 4. (a) The top most confident head detections (shown in red) used to estimate the camera parameters (height and tilt). The camera parameters are in turn used to constrain the heights of detected heads in the scene. (b) Visualization of the metric rectification of a frame of a crowd video using the estimated scene geometry.

over a set of geometrically plausible scales, a detection hypothesis for a given location in the frame is compared the reported height of the head detection in real-world coordinates to that of the average person head. When the absolute difference between the head heights of the detection hypothesis and the average head size is too large, the detection hypothesis is suppressed.

3.3. Crowd density estimation

Here we outline the density estimator $D(p)$ used to drive the density-informed detection given by cost function (1). Crowd density estimation has been addressed in a number of works [5, 6, 18, 20, 24]. While some of this work is based on counting detected people, here we seek an alternative approach to complement our person detector. We therefore resort to regression-based density estimation and follow the recent method of [18]. We assume a set of training images with dense feature map $\phi(p) \in \mathcal{R}^m$ at each pixel p and ground-truth annotations of head positions ξ_i. We define ground truth density $F^0(p)$ as a kernel density estimate based on positions of annotated points as

$$F_0(p) = \frac{1}{2\pi\tau^2} \sum_i \exp\left(- \frac{\|\xi_i - p\|^2}{2\tau^2} \right),$$

where we use a 2D Gaussian kernel with a small bandwidth $\tau = 4$ corresponding to the size of heads in our feature map ϕ. Following [18], we learn a linear transformation of feature responses that approximates the density function at each pixel $F(p|w) = w^\top \phi(p)$ where $w \in \mathcal{R}^m$ is a weight vector. We learn w from k training samples by minimizing the regularized MESA distance $\mathcal{D}$ introduced in [18] as

$$\hat{w} = \underset{w \in \mathcal{R}^m}{\operatorname{argmin}} \left(w^\top w + \lambda \sum_k \mathcal{D}(F_k^0(\cdot), F_k(\cdot|w)) \right). \tag{4}$$

The density estimator above can be used with different types of feature maps ϕ. In this paper we define ϕ using the dense scores $s(p)$ of the person detector described in Section 3.2. In particular, we define the feature map $\phi_i(p), i = 1...m$ by smoothing detector scores $s(p)$ that are

obtained from a range of geometrically plausible scales with m Gaussian filters $G(\cdot;\, t_i)$ of different size t_i.

While $F(p)$ provides per-pixel estimate of the density function, here we aggregate it in Gaussian windows of size σ and define $D(p)$ in (1) as $D(p) = F(p) * G(\cdot;\, \sigma)$. As mentioned earlier, σ is an important parameter related to the quality of the density estimator, small values of σ result in stronger density prior in our framework, but also have less reliable results of density estimation. To balance this trade-off, we choose σ by cross-validation and fix its value to 90 pixels in our experiments. We note that adapting the value of σ to the estimated size of people in the image may provide further improvements.

3.4. Tracking detections

In this section we describe how the person detections, obtained in individual frames by minimizing the density-aware cost function (1), are associated into person-tracks over multiple video frames. The objective here is to to associate head detections in individual frames into a set of head tracks corresponding to the same person within the crowd across time, thereby generating tracks of individuals that conform to the crowd. We follow the tracking by detection approach of [10], which demonstrated excellent performance in tracking faces in TV footage, but here apply it to track heads in crowded video scenes. The method uses local point tracks throughout the video to associate detections of the same person obtained in individual frames. For each crowd video sequence, we obtain point tracks using the Kanade-Lucas-Tomasi tracker [28]. The point tracks are used to establish correspondence between pairs of heads that have been detected within the crowd.

For a given pair of head detections in different frames, the number of point tracks which pass through both heads is counted, and if this number is large relative to the number of point tracks which are not in common to both heads, a match is declared. We use a single-link agglomerative grouping strategy as in [10] which gradually merges head detections into larger groups starting from the closest (most connected) detections.

This simple tracking procedure is extremely robust and can establish matches between head detections where the head has not been continuously detected due to pose variation or partial occlusions due to other members of the crowd. In Section 4 we demonstrate the improvement in detection performance using this type of tracking by association: missing detection below detection threshold can be filled-in, and short tracks corresponding to false positive detections can be discarded.

4. Evaluation

This section evaluates results of our method on the task of detecting and tracking people in crowded video scenes.

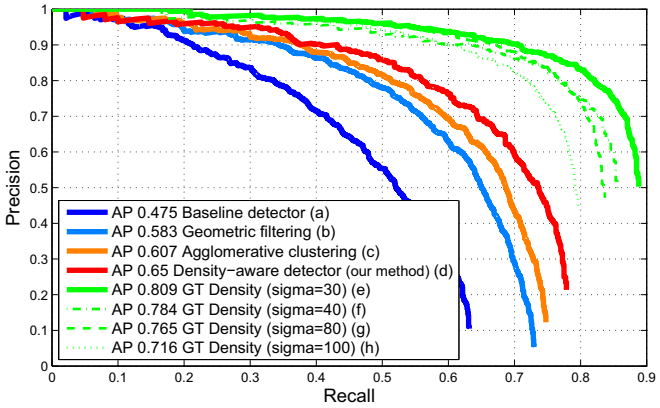


Figure 6. Evaluation of person detection performance. Precision-recall curves for the (a) baseline detector, (b) after geometric filtering, (c) tracking by agglomerative clustering and (d) using the proposed density-aware person detector. Note the significant improvement in detection performance obtained by the density-aware detector. For comparison, the plot also shows performance of the density-aware detector using the ground truth density, obtained by smoothing ground truth detections by a Gaussian with different sigmas. Note the improvement in performance for smaller sigmas. In this case, for sigma approaching zero, the density would approach the ground truth and hence the perfect performance.

Dataset. For the training and testing we have collected videos of crowd scenes from the Internet and recorded our own set of videos with scenes of political rally. These videos contain considerable variation in terms of viewing angle, scale, crowd motion and crowd density. The camera remained static in all the videos. For the training and evaluation of detection accuracy, we have annotated heads of all visible people in randomly selected frames. All the videos were divided into three subsets. The first subset with 1200 annotated head bounding boxes was used to train a person detector [12] (Section 3.2). The second subset with all people annotated in 60 frames was used to train a density estimator [18] (Section 3.3). The last test subset with 1009 annotated head bounding boxes was deployed to evaluate the performance of the detection.

In order to evaluate the task of tracking, we employed a subset of 13 video clips captured at a large political rally. Examples of the video frames from this tracking testing set are depicted in Figures 1 and 8. On average, each video clip is roughly two minutes long with a frame size of 720×480. As can be seen, this collection of crowd video clips spans a wide range of densities, viewpoints and zoom levels. We have annotated 122 person tracks to evaluate the tracking performance on videos from the test subset. This data is available on the project website [1].

Detection. To test and compare the detection performance, we follow PASCAL VOC evaluation protocol [11]. A predicted bounding box is considered correct if it overlaps more than 50% with a ground-truth bounding box, only

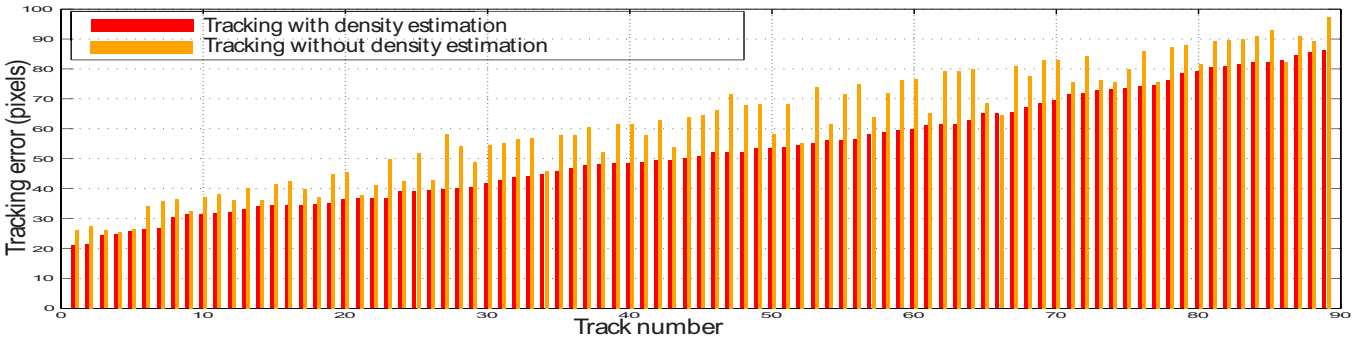


Figure 5. Tracking RMS error (in pixels) for different tracks with and without using the proposed density-aware detector. The error is measured with respect to manually labelled tracks. Note that in many cases the error is significantly reduced using the proposed method.

one detection overlapping with ground-truth bounding box is considered correct, other overlapping detections are declared as false positives. The Performance is measured in terms of precision-recall and average precision (AP) values.

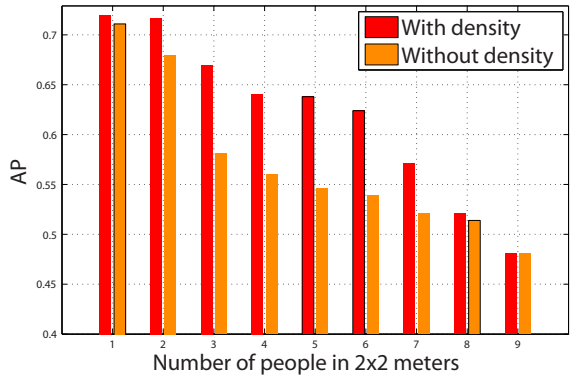


Figure 7. Person detection performance, measured by average precision (y-axis) for different person densities (x-axis). Note that proposed density-aware detector outperforms the baseline detector across almost all density-levels, and specially medium-level densities of 2-7 people in 2x2 meters.

To demonstrate the advantage of our method on the detection task, we have compared it to three alternative detectors. Our first *baseline detector* is [12] trained on our training data. The second detector integrates the baseline detector with geometric filtering imposing the prior on the size of detections as described in Section 3.2. The third detector further integrates temporal consistency constraints using agglomerative clustering described in Section 3.4. Finally, our density-aware detector optimizes the introduced cost function (1) and integrates geometric filtering and temporal consistency constraints as in the case of other detectors. The comparative evaluation is shown in Figure 6. As can be observed, the density-aware detector (red curve) outperforms all three other detectors by a large margin. Qualitative detection results are illustrated in Figure 8.

To gain understanding of the density prior introduced in this work, Figure 6 also shows detection results for the density-aware detector using ground truth density estimates obtained by smoothing ground truth person detections with Gaussians of varying σ (green curves). Interestingly, the detection performance increases significantly in this case, suggesting that our detector can further benefit from future better performing density estimation methods. As expected, the performance of the detector increases for better localized ground truth density estimators with smaller values of σ.

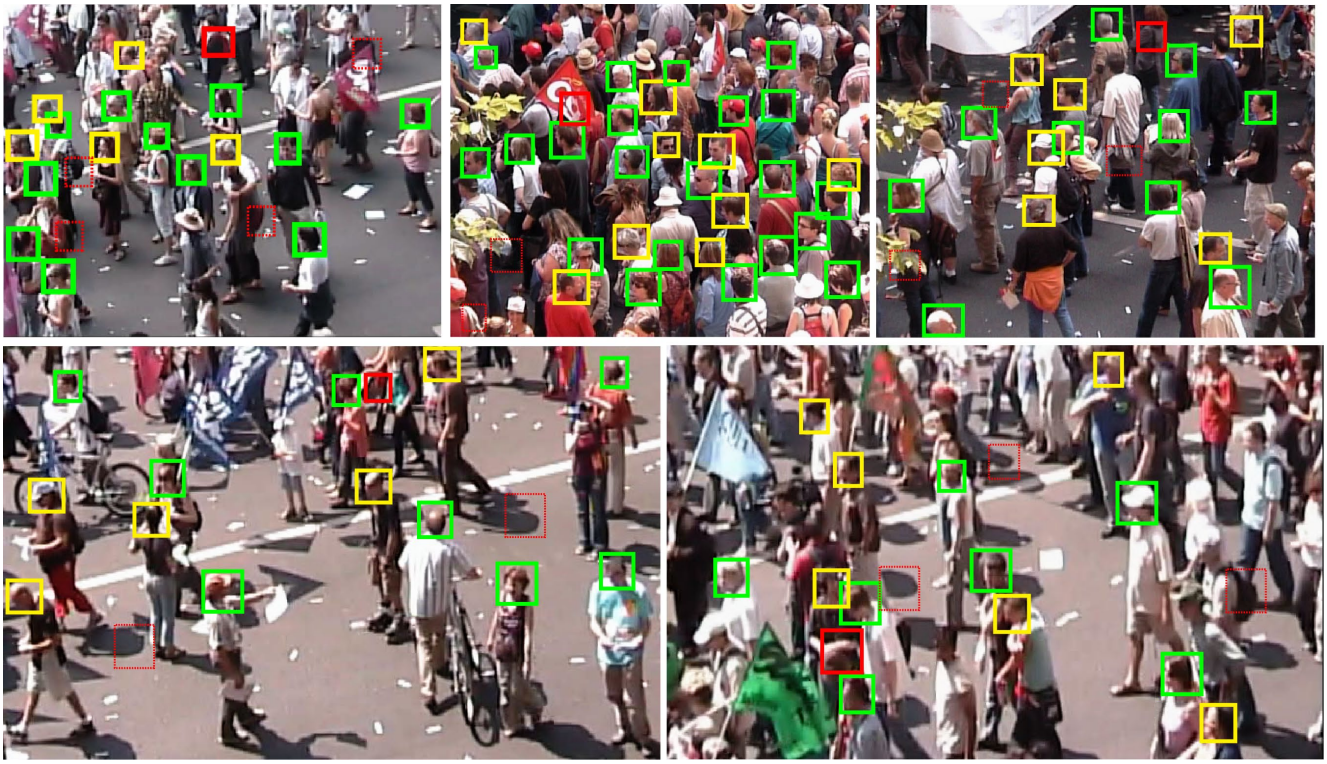
We finally illustrate detection results over scene regions with different person densities in Figure 7. We observe that the density-aware detector (red bars) obtains highest gain for the medium range of density values (2-7 people per 4m^2). This can be explained by the increasing error of the density estimator for low-density scenes, and decreasing performance of the person detector for the very high density scenes.

Tracking. The objective of this set of experiments is to assess the improvement that can be attained in tracking accuracy using the proposed density-aware crowd model in the presence of a range of crowd densities.

Quantitative analysis of the proposed tracking algorithm was performed by generating ground-truth trajectories for 122 people, which were selected randomly from the set of all people in the crowd. The ground-truth was generated by manually tracking the centroid of each selected person across the video. In our experiments we evaluate tracks independently, by measuring tracking error (measured in pixels) which is achieved by comparing the tracker position at each frame with respect to the position indicated by the ground truth. When our system does not detect a person that has been labeled in the ground truth, this corresponding track is considered lost. In total, our system was able to detect and track 89 out of the 122 labelled individuals.

A set of trajectories generated by our tracking algorithm is shown in Figure 8. The average tracking error for the annotated heads that were detected obtained using the proposed model was 52.61 pixels (Figure 5). In order to assess the contribution of density estimation in tracking accuracy, a baseline tracking procedure consisting of detection,

Detection results



Tracking results

Figure 8. Examples of detection and tracking results for different crowded scenes and levels of person density. For detection (top), true positives are shown in green, false positives are shown in red. The additional true positives detected by the proposed density-aware detector (compared to the baseline with geometric filtering and tracking) are shown in yellow. False positives removed by the proposed method are shown by red dashed line. For tracking results (bottom), head detections are depicted with green bounding boxes along with their associated tracks over 100 frames. Additional results and videos are available on the project web page [1].

geometric filtering and agglomerative clustering was evaluated. The mean tracking error of this baseline algorithm was 64.64 pixels. This increase in tracking accuracy can be attributed to the fact the association of individual head detections into tracks contains fewer false positives.

We further evaluated the ability to track people over a span of frames by measuring the difference in the length of the generated tracks in relation to the manually annotated tracks. The mean absolute difference between the length of the ground truth tracks and the tracks generated by our system was 18.31 frames, whereas the baseline (which does not incorporate density information) resulted in a mean difference of 30.48 frames. It can be observed from these results that our tracking was very accurate, in most cases, and able to maintain correct track labels over time.

5. Conclusion

We have shown that automatically obtained person density estimates can be used to improve person localization and tracking performance in crowded scenes. We have formulated the person detection task as a minimization of a joint energy function incorporating scores of individual detections, pair-wise non-overlap constraints, and constraints imposed by the estimated person density over the scene. We have demonstrated significant gains in detection and tracking performance on challenging videos of crowded scenes with varying density. Currently, video frames are processed individually and obtained detections are tracked in post-processing. Next, we plan to extend the proposed cost function to spatio-temporal video volume to jointly detect and track people in multiple frames.

Acknowledgements. This work was partly supported by the Quaero, OSEO, MSR-INRIA, ANR DETECT (ANR-09-JCJC-0027-01) and the CROWDCHECKER project. We thank Pierre Bernas, Philippe Drabczuk, and Guillaume Ne from Evitech for the helpful discussions and the testing videos.

References

[1] http://www.di.ens.fr/willow/research/crowddensity/.

[2] S. Ali and M. Shah. Floor Fields for Tracking in High Density Crowd Scenes. In *ECCV*, 2008.

[3] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transforms. In *CVPR*, 2010.

[4] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009.

[5] A. Chan, Z. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008.

[6] S. Cho, T. Chow, and C. Leung. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, 29(4):535–541, 2002.

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

[8] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

[9] X. Descombes, R. Minlos, and E. Zhizhina. Object extraction using a stochastic birth-and-death dynamics in continuum. *Journal of Mathematical Imaging and Vision*, 33(3):347–359, 2009.

[10] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... buffy–automatic naming of characters in tv video. In *BMVC*, 2006.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[12] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 32(9), 2010.

[13] A. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2000.

[14] D. Hoiem, A. Efros, and M. Hebert. Putting Objects in Perspective. *IJCV*, 2008.

[15] A. Johansson, D. Helbing, and P. Shukla. Specification of the social force pedestrian model by evolutionary adjustment to video tracking data. *Advances in Complex Systems*, 10(2):271–288, 2007.

[16] Z. Khan, T. Balch, and F. Dellaert. MCMC-based particle filtering for tracking a variable number of interacting targets. *IEEE PAMI*, 27(11):1805–1819, 2005.

[17] D. Kong, D. Gray, and H. Tao. A viewpoint invariant approach for crowd counting. In *ICPR*, 2006.

[18] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *NIPS*, 2010.

[19] K. Li and T. Kanade. Cell population tracking and lineage construction using multiple-model dynamics filters and spatiotemporal optimization. In *IWMIAAB*, 2007.

[20] A. Marana, S. Velastin, L. Costa, and R. Lotufo. Estimation of crowd density using image processing. In *IEE Colloquium on Image Processing for Security Applications*, 2002.

[21] E. Marieb and K. Hoehn. *Human anatomy & physiology*. Pearson Education, 2007.

[22] P. Nillius, J. Sullivan, and S. Carlsson. Multi-target tracking: Linking identities using bayesian network inference. In *CVPR*, pages II:2187–2194, 2006.

[23] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2010.

[24] V. Rabaud and S. Belongie. Counting crowded moving objects. In *CVPR*, 2006.

[25] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *CVPR*, 2007.

[26] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *ICCV*, 2009.

[27] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. A. Data-driven crowd analysis in videos. In *ICCV*, 2011.

[28] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994.

[29] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2), 2003.

[30] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, 2005.

[31] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.