Small Instance Detection by Integer Programming on Object Density Maps

Zheng Ma, Lei Yu, Antoni B. Chan
Department of Computer Science, City University of Hong Kong.

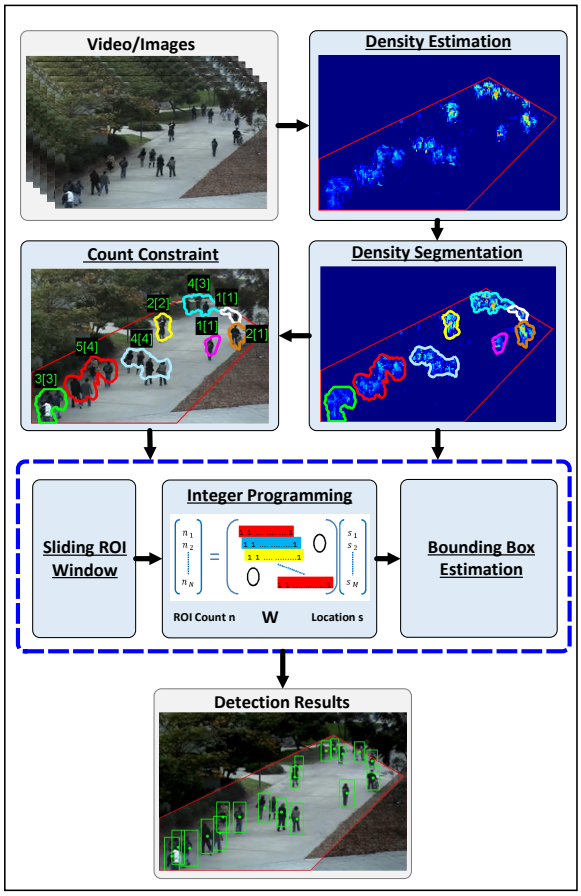

Figure 1: The proposed small-instance object detection framework. The object density map is estimated from the input image or video frame. The density map is divided into several local regions. For each region, the global count is calculated. Then, integer programming with a global count constraint is used to recover the locations of each object from the density map. Finally, the object bounding boxes are estimated from the density map.

Motivation: Recently, [2] achieves state-of-the-art counting performance by estimating an object density map from an input image. The object density map indicates the distribution of the objects within the image, and integrating over an ROI in the density maps yields an estimate of the object count within the ROI.

Inspired by [2], we propose a novel detection framework for small-instances using object density maps (see Fig. 1), which takes full advantage of counting methods and avoids the potential drawbacks of using traditional detection methods on small-instances. This is a new joint counting and detection framework, which can output both counting and detection results. The contributions of this paper are four-fold. First, we develop a 2D integer programming method that recovers 2D object locations from an object density map. Our framework is unique in that there is no non-maximum suppression or detector threshold to select. Second, to take full advantage of the accurate counting results from [2], a global count constraint is added to the integer programming objective function. The constraint regularizes the detection results by suppressing false positive errors and increasing recall when there is heavy occlusion. Third, given a detected object location, we propose a method to estimate the bounding box of the object using the density map. Last, the proposed detection method achieves state-of-the-art results on several quite different and challenging small-object datasets, including pedestrians, cells, fish, flies, honeybees, and seagulls.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

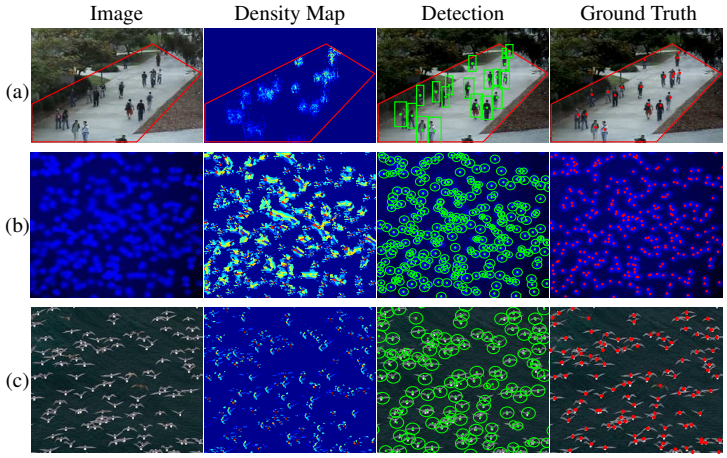

Figure 2: Results of the proposed detection method on (a) UCSD pedestrians, (b) Synthetic cells, (c) Seagulls. The red dots are the ground truth locations and the green boxes/circles are the detection results.

Density Map Estimation: To estimate the object density map, we use the method from [2], which learns the mapping from the extracted image feature to the density value for each pixel, $F(p;w) = w^T x_p$, where w is the weight matrix and x_p is the feature vector at pixel p.

Localization by Integer Programming: We propose a 2D integer programming method to recover the locations of objects in the estimated density map. Given the input density map $F(p;w)$ for an image $\mathcal{I}$, the objective is to find the object indicator map $S(p), \forall p \in \mathcal{I}$, where the value $S(p)$ indicates the number of objects present at location p (e.g., 0 means there is no object at location p, and 1 means there is 1 object at p). Let the vector $\mathbf{s} \in Z_+^{|\mathcal{I}|}$ be the vectorized matrix S, and $\mathbf{f}$ be the vectorized density map F.

Next, we define a set of N sliding windows over the density map, where the window size is set as the average object size. Each window is represented as a mask vector $\mathbf{w}_i \in \{0,1\}^{|\mathcal{I}|}$, where the entries are 1 for pixels in the ROI of the window, and 0 otherwise.

Using the density map, the number of objects in the sliding window $\mathbf{w}_i$ is estimated as $n_i \approx \mathbf{w}_i^T \mathbf{f}$. On the other hand, the number of objects in the same window according to the object indicator map is $n_i = \mathbf{w}_i^T \mathbf{s}$. Hence the optimal $\mathbf{s}$ can be obtained by minimizing the L1-norm between these two equations. We also add a global count constraint term to ensure that the total number of detected objects $n_c = \mathbf{1}^T \mathbf{s}$ is close to the number predicted by the density map $n_c \approx \mathbf{1}^T \mathbf{f}$, where $\mathbf{1}$ is the vector of ones. Hence,

$$\mathbf{s}^* = \underset{\mathbf{s} \in Z_+^{|\mathcal{I}|}}{\operatorname{argmin}} \sum_{i=1}^{N} |\mathbf{w}_i^T \mathbf{s} - n_i| + \lambda |\mathbf{1}^T \mathbf{s} - n_c|, \tag{1}$$

where λ is the regularization parameter. Note that n_i and n_c are calculated from the density map and $\mathbf{w}_i$ is fixed for an image, and hence finding $\mathbf{s}$ is a signal reconstruction problem with non-negative integer constraints on its entries. The formulation in (1) is a linear integer programming problem, which we solve using CPLEX [1].

Bounding Box Estimation: For each detected object location, the optimal bounding box is found such that the integral of the density map under the box is close to 1.

Results: Examples of the detection results are shown in Fig. 2.

[1] IBM. Ibm ilog cplex optimizer. http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/, 2013.

[2] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, 2010.