Single-Image Crowd Counting via Multi-Column Convolutional Neural Network

Liangjie Cao
8 June 2018

Abstract

The paper aims to develop a method than can accurately estimate the crowd count from an individual image with arbitrary crowd density and arbitrary perspective. To this end, the authors have proposed a simple but effective Multi-column Convolutional Neural Network (MCNN) architecture to map the image to its crowd density map. The proposed MCNN allows the input image to be of arbitrary size or resolution. By utilizing filters with receptive fields of different sizes, the features learned by each column CNN are adaptive to variations in people/head size due to perspective effect or image resolution. Furthermore, the true density map is computed accurately based on geometry-adaptive kernels which do not need knowing the perspective map of the input image. Since exiting crowd counting datasets do not adequately cover all the challenging situations considered in our work, the authors have collected and labelled a large new dataset that includes 1198 images with about 330,000 heads annotated. On this challenging new dataset, as well as all existing datasets, they conduct extensive experiments to verify the effectiveness of the proposed model and method. In particular, with the proposed simple MCNN model, their method outperforms all existing methods. In addition, experiments show that their model, once trained on one dataset, can be readily transferred to a new dataset.

1. Introduction

Many algorithms have been proposed in the literature for crowd counting. Earlier methods [8] adopt a detection-style framework that scans a detector over two consecutive frames of a video sequence to estimate the number of pedestrians, based on boosting appearance and motion features. [7, 9, 10] have used a similar detection-based framework for pedestrian counting. In detection-based crowd counting methods, people typically assume a crowd is composed of individual entities which can be detected by some given detectors [3, 11, 6, 2]. The limitation of such detection-based methods is that occlusion among people in a clustered environment or in a very dense crowd significantly affects the performance of the detector hence the final estimation accuracy.

In this paper, they aim to conduct accurate crowd counting from an arbitrary still image, with an arbitrary camera perspective and crowd density (see Figure 1 for some typical examples). To overcome many challenges, in this work, they propose a novel framework based on convolutional neural network (CNN)[4] for crowd counting in an arbitrary still image. More specifically, they propose a multi-column convolutional neural network (MCNN) inspired by the work of [1], which has proposed multi-column deep neural networks for image classification. In their model, an arbitrary number of columns can be trained on inputs preprocessed in different ways. Then final predictions are obtained by averaging individual predictions of all deep neural networks. Their MCNN contains three columns of convolutional neural networks whose filters have different sizes. Input of the MCNN is the image, and its output is a crowd density map whose integral gives the overall crowd count.

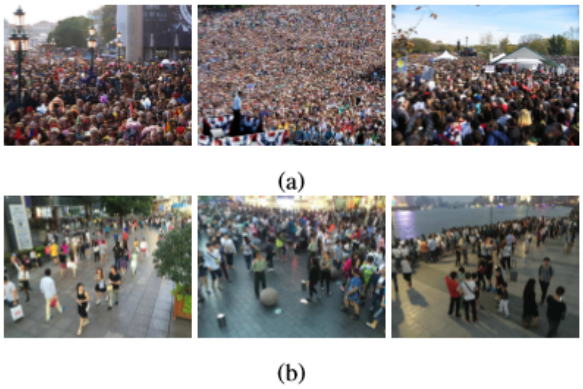

Figure 1. **(a) Representative images of Part A in their new crowd dataset. (b) Representative images of Part B in their crowd dataset. All faces are blurred in (b) for privacy preservation.**

2. Multi-column CNN for Crowd Counting

To estimate the number of people in a given image via the Convolutional Neural Networks (CNNs), there are two natural configurations. One is a network whose input is the image and the output is the estimated head count. The other

one is to output a density map of the crowd (say how many people per square meter), and then obtain the head count by integration.

3. Density map via geometry-adaptive kernels

Since the CNN needs to be trained to estimate the crowd density map from an input image, the quality of density given in the training data very much determines the performance of our method. They first describe how to convert an image with labeled people heads to a map of crowd density. If there is a head at pixel xi, they represent it as a delta function $\delta(x - x_i)$. Hence an image with N heads labeled can be represented as a function:

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i) \quad (1)$$

To convert this to a continuous density function, they may convolve this function with a Gaussian kernel [5] G so that the density is F(x) = $H(x) * G_\sigma(x)$. However, such a density function assumes that these x_i are independent samples in the image plane which is not the case here: In fact, each x_i is a sample of the crowd density on the ground in the 3D scene and due to the perspective distortion, and the pixels associated with different samples x_i correspond to areas of different sizes in the scene.
Therefore, they should determine the spread parameter based on the size of the head for each person within the image. However, in practice, it is almost impossible to accurately get the size of head due to the occlusion in many cases, and it is also difficult to find the underlying relationship between the head size the density map. Interesting they found that usually the head size is related to the distance between the centers of two neighboring persons in crowded scenes (refer to Figure 2). As a compromise, for the density maps of those crowded scenes, they propose to dataadaptively determine the spread parameter for each person based on its average distance to its neighbors.

Figure 2. **Original images and corresponding crowd density maps obtained by convolving geometry-adaptive Gaussian kernels**

References

[1] *D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In* CVPR*, 2012.*

[2] *W. Ge and R. T. Collins. Marked point processes for crowd counting. In* CVPR*, 2009.*

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

[4] *Y. Lcun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition.* P IEEE*, 1998.*

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

[6] *M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In* ICPR*, 2008.*

[7] *Z. Lin and L. S. Davis. Shape-based human detection and segmentation via hierarchical part-template matching.* IEEE TPAMI*, 2010.*

[8] *P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance.* IJCV*, 2005.*

[9] *M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In* CVPR*, 2011.*

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

[11] *C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In* CVPR*, 2008.*