J. Vis. Commun. Image R. 38 (2016) 530–539

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Dense crowd counting from still images with convolutional neural networks ☆

Yaocong Hu, Huan Chang, Fudong Nian, Yan Wang, Teng Li *

Anhui University, No. 111 Jiulong RD, Hefei 230061, China

ARTICLE INFO

Article history:
Received 1 July 2015
Revised 29 February 2016
Accepted 20 March 2016
Available online 29 March 2016

Keywords:
Crowd counting
Convolutional neural networks
Feature learning
Regression

ABSTRACT

For reasons of public security, modeling large crowd distributions for counting or density estimation has attracted significant research interests in recent years. Existing crowd counting algorithms rely on predefined features and regression to estimate the crowd size. However, most of them are constrained by such limitations: (1) they can handle crowds with a few tens individuals, but for crowds of hundreds or thousands, they can only be used to estimate the crowd density rather than the crowd count; (2) they usually rely on temporal sequence in crowd videos which is not applicable to still images. Addressing these problems, in this paper, we investigate the use of a deep-learning approach to estimate the number of individuals presented in a mid-level or high-level crowd visible in a single image. Firstly, a ConvNet structure is used to extract crowd features. Then two supervisory signals, i.e., crowd count and crowd density, are employed to learn crowd features and estimate the specific counting. We test our approach on a dataset containing 107 crowd images with 45,000 annotated humans inside, and each with head counts ranging from 58 to 2201. The efficacy of the proposed approach is demonstrated in extensive experiments by quantifying the counting performance through multiple evaluation criteria.

1. Introduction

Crowd counting aims at calculating the number of individuals presented in images and videos. It is an important topic with many potential practical applications, such as video surveillance (e.g., dense crowd anomaly detection, crowd management in a specific region), safety management (e.g., recording the number of people entering or leaving some regions) and web multimedia (e.g., estimating crowd size of tweet image shoot in crowd scenic spot). However, due to problems including crowd variation, occlusion, clutter and low resolution, visual analysis based crowd counting and density estimation are still very challenging tasks.

The task of crowd counting has been approached from a number of angles, but the techniques share a common framework: crowd feature extraction, followed by crowd counting using object detection or regression model. However, crowds can be various in their distributions and color patterns. And more importantly, a crowd does not have a well-defined shape as a single object does, which makes it difficult for crowd feature extraction. These difficulties cause that existing methods well-suited in pedestrian detection cannot be applicable in detecting human instances in crowd scenes. For example, we apply Deformable Parts Model (DPM) [1] in Fig. 1. Detection results show that this detection-based method is more applicable in the crowd of few tens than in the crowd of more than hundreds.

To address these challenges, some research works [2–4] indicate that the crowd in high density scenes often presents repetitive textural visual effects, namely, the crowd distribution is irregular and nonuniform in large scales, but it presents some regular visual patterns in small scales. Moreover, in derived intensity spaces such as image derivative, or edges, groups of individuals are likely to exhibit an increased level of similarity [3]. For reasons stated above, how to extract features that can well represent the information contained in the crowd is especially vital for the following procedure of this task.

In recent years, with the success of deep learning architectures for visual processing, (e.g., convolutional neural networks (ConvNets)) and availability of image databases with millions of labeled examples (e.g., ImageNet) [5], the state of the art in many different domains are advancing rapidly, including image classification [5,6], object and face detection [7,8], speech recognition [9], bioacoustics [10], etc. Unlike many previous vision approaches using hand-designed features, ConvNets can automatically learn a unique set of features optimized for a given task. Recent researches

☆ This paper has been recommended for acceptance by Dacheng Tao.

* Corresponding author.
E-mail address: liteng@ahu.edu.cn (T. Li).

http://dx.doi.org/10.1016/j.jvcir.2016.03.021
1047-3203/© 2016 Elsevier Inc. All rights reserved.

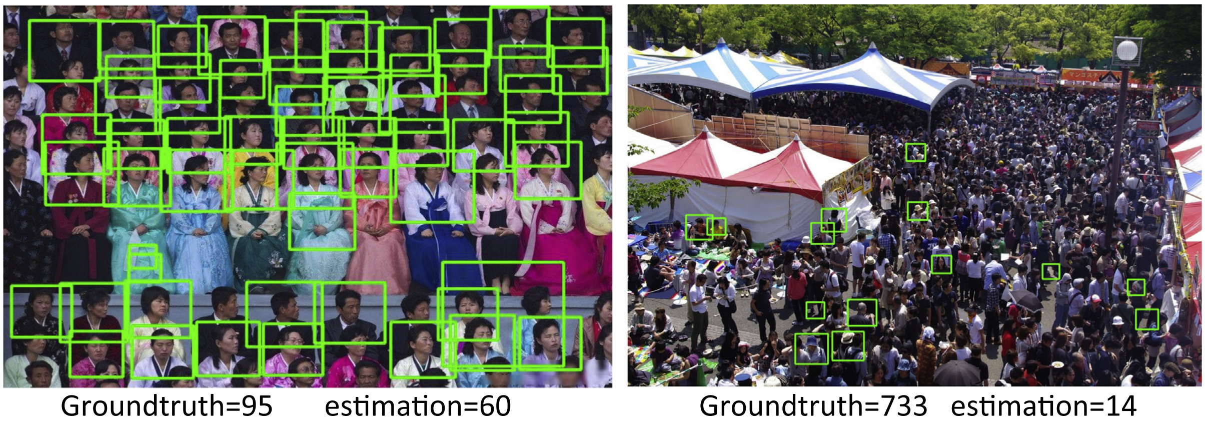

Fig. 1. Some typical examples of human detection.

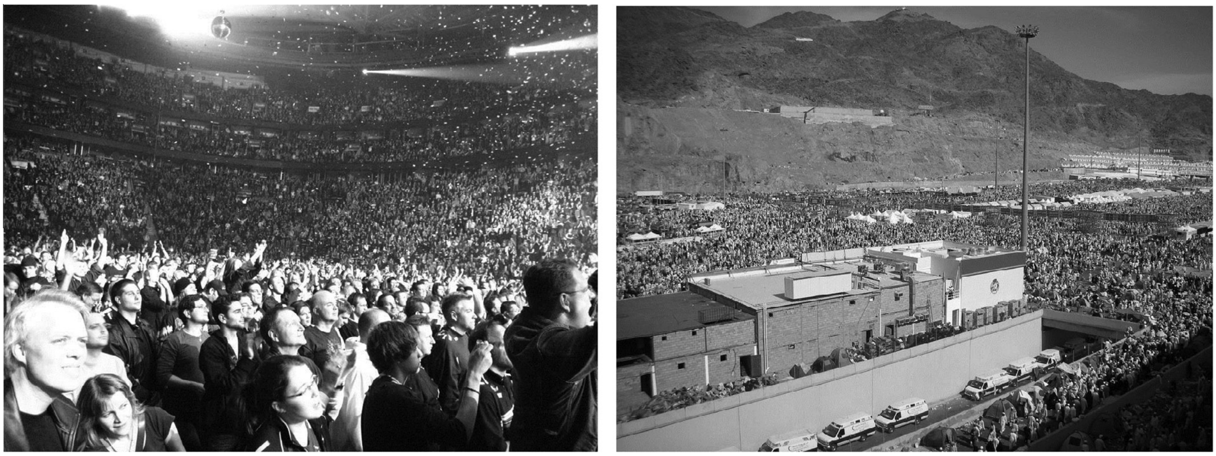

Fig. 2. Two examples from UCF-CROWD dataset [3]. In both images, people far away from the scene only occupy so few pixels that even a human observer cannot distinguish them from the background.

also have shown that learned features are able to perform better than traditional hand-engineered representations (e.g., Scale Invariant Feature Transform (SIFT) [11], Histogram of Oriented Gradient (HOG) [12], Local Binary Patterns (LBP) [13]) in many domains, especially those where good features have not already been engineered [14].

Inspired by the effective and superior features learned with the deep architecture, this paper develops a simple and general discriminative learning-based framework for the problem of people counting in images. Firstly, a ConvNet structure is used to learn crowd features and then, a feature-count regressor considering two supervisory signals, i.e., crowd count and crowd density, maps the learned feature to the number of people within each local region. As a result, the total crowd estimation is the sum of that in all local regions. The proposed method evades the traditional task of learning to detect and localize individual object instances in mid-level and high-level crowd density images, which is impractical in many cases. Our sole aim is to use feature vectors learned in ConvNets to estimate people count in each local region and we expect the deviation between our estimation and groundtruth can be as small as possible.

In terms of experimental datasets, most of the previous crowd counting algorithms only have been verified on low density crowd datasets, e.g., USCD dataset [15,16] with people count of 11–46, Mall dataset [17] with count of 13–53 individuals and PETS dataset [18] containing 3–40 people per frame. To the best of our knowledge, so far, only Idrees et al. [3] provided their UCF-CROWD dataset containing between 94 and 4543 people per image, and their crowd counting algorithm achieved state-of-the-art performance on these dense crowd images. However, in some extreme dense crowd images of UCF-CROWD dataset, an individual only occupies so few pixels that even a human observer cannot distinguish it from background (as shown in Fig. 2), and such images actually are of no practical use in real world applications. To address this problem, in this paper, we provide our own AHU-CROWD dataset covering different scenarios, with head counts from 58 to 2201 per image, and all individuals visible in our dataset can be well distinguished by human observers. Moreover, in accordance with paper [3,19], images are annotated with dots, which is the natural way to count objects for humans, at least when the number of objects is large. Fig. 3 gives some examples of the counting problems and the dotted annotations we consider in this paper. The main contributions of our study can be concluded into three aspects:

- We propose a deep learning architecture to estimate the people counting in still images.
- We provide our own crowd dataset AHU-CROWD, which consists of 107 crowd images covering different scenes. All 45 K human instances are annotated with dots manually (one dot per person).