Deep People Counting in Extremely Dense Crowds

Chuan Wang[1], Hua Zhang[1], Liang Yang[1,2], Si Liu[1], Xiaochun Cao[1*]
[1]State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences.
[2]School of Information Engineering, Tianjin University of Commerce.
{wangchuan, zhanghua, yangliang, liusi, caoxiaochun}@iie.ac.cn

ABSTRACT

People counting in extremely dense crowds is an important step for video surveillance and anomaly warning. The problem becomes especially more challenging due to the lack of training samples, severe occlusions, cluttered scenes and variation of perspective. Existing methods either resort to auxiliary human and face detectors or surrogate by estimating the density of crowds. Most of them rely on hand-crafted features, such as SIFT, HOG *etc*, and thus are prone to fail when density grows or the training sample is scarce. In this paper we propose an end-to-end deep convolutional neural networks (CNN) regression model for counting people of images in extremely dense crowds. Our method has following characteristics. Firstly, it is a deep model built on CNN to automatically learn effective features for counting. Besides, to weaken influence of background like buildings and trees, we purposely enrich the training data with expanded negative samples whose ground truth counting is set as zero. With these negative samples, the robustness can be enhanced. Extensive experimental results show that our method achieves superior performance than the state-of-the-arts in term of the mean and variance of absolute difference.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

People counting; convolutional neural networks(CNN); crowd analysis;

1. INTRODUCTION

Crowd analysis, *e.g.*, estimating the number of people in the crowd, is becoming one of the most important and challenging problems in the field of multimedia and computer vision. The recurrent tragic stampedes at New Year's Eve or pilgrimages and terrorist attacks at the marathon or squares have shown great significance of crowd analysis on abnormal states alarm. Hence we urgently require powerful and sophisticated approaches for visual analysis of dense crowds. Besides, crowd analysis is useful in arrangement, modification and expansion of traffic facilities (tunnels, overpasses and traffic lights), as well. It can also be used to detect, track and manage crowd events like marathons, protests, and rallies that all may gather hundreds of people. A lot of works on crowd analysis are proposed including crowd detection [1], crowd segmentation [7], and collective motion learning for anomaly detection.

Figure 1: Extremely dense crowd samples.

In recent years, great efforts have been made for counting people in natural scenes [2, 3], which can be divided into two categories: direct and indirect estimation. Direct methods are mostly based on human detection achieved by face or parts filters of people [2]. Indirect ones always obtain the number of people by estimating crowd density ranges. The general strategy is to learn a regression model between density and low level features belonging to corresponding region. Although the traditional methods have achieved much success, most of them verifies on low or medium density crowds. However, in particular scenes like concerts, rallies and protests, thousands of people may exist and each person may merely be represented by no more than 10 of pixels in photo records. These situation will certainly make traditional methods failed. With increase of density, decrease of number of pixels each person occupies or existence of severe occlusion among people, crowd counting becomes more difficult and most of existing techniques may degrade

*Corresponding author

MM'15, October 26-30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.28063370-12345-67-8/90/01.

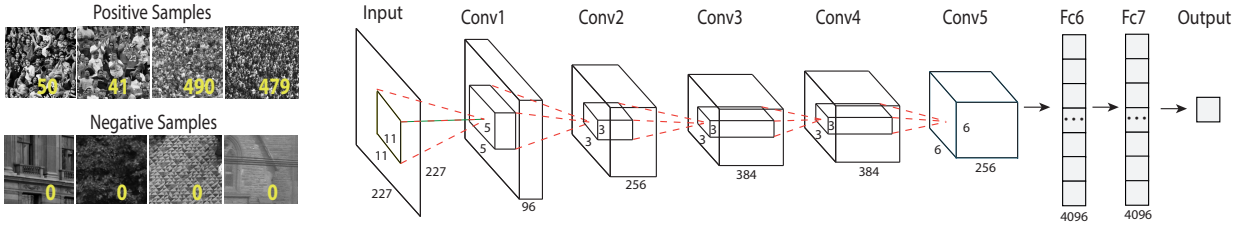


Figure 2: Deep model architecture. We feed this deep network with the patches of crowd images and negative sample images, and its output is the estimated people counts in the input patch.

or fail. Moreover, traditional methods rely on hand-crafted features, such as SIFT and HOG, which are turned to be sub-optimal for this task. Since cluttered scene, perspective distortion, or even environment factors like illumination can provide misleading information and lead traditional features based human detection methods, *e.g.*, Deformable Part Models (DPMs), invalid and prevent people counting from making success.

Recently, deep models especially convolutional neural networks have been widely applied to many computer vision problems like image classification, object detection and face recognition [9]. Girshick *et al.* [4] points out that features extracted from convolutional neural networks [8] trained for classification are sometimes more effective than hand-crafted features and can be reusable to broader tasks. Effectiveness of leveraging convolutional neural networks in people counting task has not been fully explored.

In this paper we develop a deep regression model for counting people in extremely dense crowd images using deep convolutional neural networks (CNN). Firstly, we adopt CNN as our basic framework to learn efficient features for counting and thus develop an end-to-end system. Secondly, we feed the CNN with expanded negative samples to reduce false alarms caused by specific factors caused by dense crowds like cluttered scenes. We collect a few images without people and set their regression score as 0, which makes our method more robust.

2. METHOD

We firstly describe the data collection and preparation in Section 2.1. Then we detail the adopted deep model in Section 2.2. Finally we present how we obtain negative samples in Section 2.3.

2.1 Data Collection and Preparation

Due to the requirement of large number of training data announced by CNN, we collect a set of images from publicly available websites, such as Google and Flickr, and mark people using a manually designed dotting tool. This set consists of 51 images each of which has 731 people on average. The counts range from 95 to 3714 as shown in Figure 3. These images cover various events including stadiums, concerts, rallies and Color Run *etc.*

Perspective effects can induce large variation of people, which means some people may merely occupy a few pixels whereas others may take a large regions separately. Consequently when cropping images with larger image size and lower density of people, operation with fixed and uniform size, which consisted with that employed on dense crowds, may result in a lot of samples with partial face or body across the full patches and hence loss representation power of cropped patches. Besides, we prefer to pay more attention on counting on extremely dense crowds in which bodies of most people may not been seen literally. Thus partial samples cropped by uniform size is definitely profitable, they may insufficient and make little effects on learning model for regression. To alleviate influence of partially cropping, we manually pick out images which may need multi-scale crop size and operate crop step with other size. Most of these pairs are cropped with fixed top left points and side length as 672 and 896, respectively. Totally we get a number of 5,705 training samples.

Firstly, we warp the cropped gray-scale patches with 227× 227 pixels. Secondly we randomly crop 224 × 224 patches at four corners and the center over input patches and horizontally flipped for data augmentation. Finally the warped augmented mean-subtracted patches with a total number of 6,414 are fed to CNN and a 1-dimensional feature vector is inferred by forward propagating.

2.2 Convolutional Neural Network for Regression

Considering the excellent performance of CNN performs in most of computer vision tasks, we employ a CNN architecture as our base framework. This architecture consists of five convolutional, some of which are followed by pooling layers, and three fully connected layers. A schematic diagram of this architecture is presented in Fig. 2.

Convolutional layer. It convolves the input images with linear sliding filters to generate response maps. If $\mathbf{X}_i$ is the feature map of input or output of i^{th} layer, this convolutional operation can be denoted as

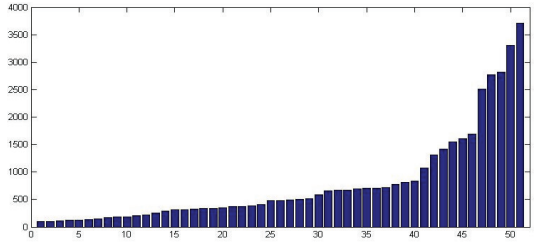
$$\mathbf{X}_i = \mathbf{W}_i \otimes [\mathbf{X}_{i-1}, 1]^T$$

where $\mathbf{W}_i = [\mathbf{W}_{i1}, \mathbf{W}_{i2}, ..., \mathbf{W}_{ik}, \mathbf{b}_i]$ indicates filters' parameters of the i^{th} layer, $\otimes$ indicates the convolution operation and $\mathbf{b}_i$ indicates the bias.

Pooling layer. Similar to Bag-of-Words (BoW) [10] and spatial pyramid [5] operated behind extracted encoded SIFT vectors, the deep convolutional features obtained by convolving can be pooled in a analogous way. Pooling layer down-samples the convolutional features non-linearly in a pattern of maximum or average over sliding-window subregions. It can significantly reduce the number of network parameters;

Fully-connected layer. It can be computed as

$$\mathbf{y}_{im} = \mathbf{W}_{im}\mathbf{X}_{(i-1)} + \mathbf{b}_m$$



(a) Statistics of collected dataset

(b) Least GroundTruth: 95

(c) Most GroundTruth: 3714

Figure 3: Statistics and samples of the collected dataset. The statistics information is shown in (a). X-axis and y-axis denote the image id and the number of annotated person in the corresponding image, respectively. (b) and (c) demonstrate the images containing the least and the most people.

where $\mathbf{W}_{im}$ denotes the m^{th} filter of the i^{th} layer. This layer requires fixed number of inputs and outputs to convert response maps as close to the ground truth;

Neuron layer. It applies nonlinear activations, such as hyperbolic tangent function or rectified linear unit, between a neuron's output and the input of next layer.

Loss layer. For regression problem, the loss function is defined as the sum of squares of differences between the ground truth and prediction,

$$LOSS = \frac{1}{2N}\sum_{j=1}^{N}\left(p^j - g^j\right)^2,$$

where p^j and g^j are the predicted and ground truth numbers of people in patch j, respectively.

It has been proved that network trained for image classification [8] can also be effective on other visual computing tasks. Thus, we replace the last fully-connected layer consisted of 4096 items with that consisted of 1 item that means the predicted number of people. Then we use the pre-trained parameters of five convolutional layers and two fully-connected layers and fine-tune the model by the following collected crowd images.

2.3 Negative Samples

Our framework is an end-to-end system which automatically obtains the high responses around the crowds instead of resorting to auxiliary human or face detectors. Since we don't employ any human appearance or localization relevant information, some specific regions without person would also get high responses, such as tall lush trees, buildings with plenty of windows, decorative patterns *etc.* These high responses cause a lot of false alarms and seriously affect the performance. The main reason may be the similar appearance between objects and these crowds.

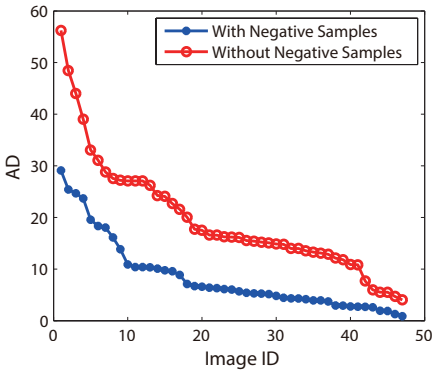


Figure 4: The impact of negative samples on performance. X-axis indicates image IDs rearranged from large AD to small. By feeding the deep model with negative samples, the mean value of AD can carry out a significant decrease.

Therefore we expand our training data by adding a number of samples with zero label containing the above mentioned factors. We add 709 negative samples, which do not contain any person, to the training set, including lush trees, buildings and some natural scene images.

Table 1: Quantitative Comparison on UCFCC.

Methods	AD	NAD
Fourier[6]	13.8±21.3	96.4±200.4
F+confidence[6]	11.0±19.7	58.7±74.9
Fc+Head[6]	11.1±19.3	63.3±84.0
FHc+SIFT[6]	10.2±18.9	53.3±69.5
traditional CNN	13.9±19.9	74.1±151.4
Our method	**8.5±15.0**	**38.9±63.1**

3. EXPERIMENTS

We use standard evaluation criteria, mean and deviation of Absolute Difference (AD) and these of Normalized Absolute Difference (NAD), to quantify the performance. NAD can be obtained by normalizing AD with the ground truth count for each patch or image. For comparison, we evaluate the performance of our method on one public people counting dataset UCFCC [6]. It consists of 50 images with a total number of 61,396 people with and count of each image ranges among 94 and 4,543. The number of average people per image could achieve 1,306 which surpass the existing crowd datasets.

In this paper we develop a deep regression model for counting people in extremely dense crowd images using deep model. To verify the influence of negative samples, we compare the performance with and without negative samples. As shown in Figure 4, by feeding CNN with negative samples, it is obvious that robustness is significantly improved with almost 50% decrease on mean value of AD. Besides, we also show mean and deviation of AD and NAD generated by CNN and our method in Table 1, and it also achieves almost 50% improvement on mean of AD and NAD by applying expanded negative samples.

Extensive experimental results show that, compared with the state-of-the-arts, our method achieves superior perfor-

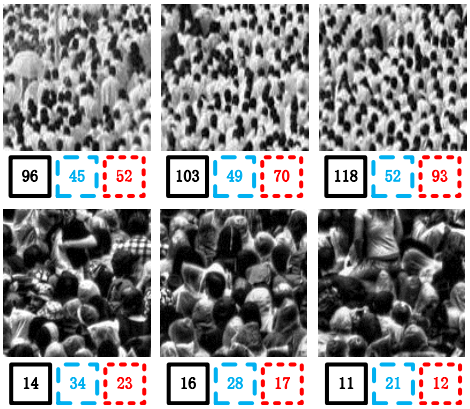
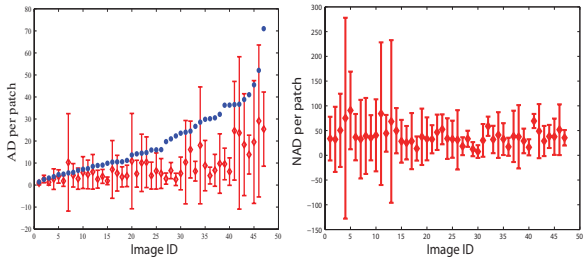


Figure 5: Quantitative comparison on 6 patches of UCFCC dataset. The numbers in black boxes are the ground truth of people in the patches, while the numbers in blue and red dashed boxes are the estimated numbers from Haroon et al. [6] and our framework respectively. It is obvious that the results of our algorithm are closer to the ground truth.

mance both in terms of mean and variance of absolute difference. For comparison, we use methods of Haroon Idrees [6] which obtain the-state-of-the-art on UCFCC dataset. We report our quantitative results both in Table 1 and Figure 5, which shows that our proposed method achieves superior performances. The first four rows in Table 1 show the improvements of gradually integrating a set of different factors including Fourier Analysis, Head detection and SIFT feature. Finally it gives best result with AD as 10.2 and NAD as 53.3. The last two rows show excellent performance of our deep model, with comprehensive large decrease of mean and deviation of AD and NAD, a descend of 16.7% and 27.0% to mean of AD and NAD, respectively. In Figure 5 we present 6 selected patches from UCFCC dataset and show improving results leading by our method. We display ground truth, estimation made by [6] and estimation made by our method in black, blue and red box, separately. Our method can deal with not only scenes available to head detection (the below row) but also the dotted ones lacking of human appearance (the upper one). Besides, we present AD and NAD of patches in Figure 6, respectively. The mean AD and NAD of patches per test image is shown with red diamond, deviation with red var, and ground truth with blue dot. It can be seen that although the ground truth increases, mean ADs change in a small region and NADs are almost stable except for ones in region 4 to 15 since less people exist in images.

4. CONCLUSION

In this paper we develop a deep regression model for counting people in extremely dense crowd images. We adopt CNN as our basic framework to learn efficient features for counting. By feeding the deep model with negative samples, the robustness is significantly improved and false alarms are remarkably suppressed. Extensive experimental results show that, compared with the state-of-the-arts, our method achieve superior performance in term of the mean and variance of absolute difference.



(a) AD an UCFCC (b) NAD on UCFCC

Figure 6: Patch estimation performance on UCFCC dataset. The x-axis shows the image IDs along with ascending ground truth counts. The means and standard deviations are shown in red asterisk and bar, while the ground truth counts in blue dot.

5. ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 61422213, 61332012,), 100 Talents Programme of The Chinese Academy of Sciences, "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA06010701), in part by the Foundation for the Young Scholars by the Tianjin University of Commerce under Grant 150113, and in part by the National Training Programs of Innovation and Entrepreneurship for Undergraduates under Grant 201410069040.

6. REFERENCES

[1] O. a. Arandjelovic. Crowd detection from still images. In *BMVC*, pages 53.1–53, 2008.

[2] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[3] W. Ge and R. Collins. Marked point processes for crowd counting. In *CVPR 2009*, pages 2913–2920, 2009.

[4] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR 2014*, pages 580–587, 2014.

[5] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV 2005*, pages 1458–1465, 2005.

[6] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR 2013*, pages 2547–2554, 2013.

[7] K. Kang and X. Wang. Fully convolutional neural networks for crowd segmentation. *CoRR*, abs/1411.4464, 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS 2012*, pages 1097–1105, 2012.

[9] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan. Matching-cnn meets KNN: quasi-parametric human parsing. *CoRR*, abs/1504.01220, 2015.

[10] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.