Modified AlexNet for Dense Crowd Counting

YOUMEI ZHANG, FALIANG CHANG, NANJUN LI, HONGBIN LIU
and ZHENDI GAI

ABSTRACT

This paper presents a modified AlexNet to estimate the number of people in still images. The sizes of people’s heads in images differ greatly due to some factors such as perspective effect and image resolution. Feature maps with different receptive fields which are adaptive to different sizes of objects are used in this crowd counting structure. We utilize parts of AlexNet to extract features, thus getting well-trained parameters to initialize the Convolutional Neural Network. Then feature maps in different convolutional layers are merged for further feature extraction. Since the feature maps are with different sizes of receptive fields, the network is more adaptive to diversity in people’s head-size. Experiments conducted on Shanghaitech dataset demonstrate the effectiveness of the proposed method.

KEYWORDS

Modified AlexNet, Crowd counting, Receptive field.

INTRODUCTION

As a content spot of crowd analysis, crowd counting has enjoyed massive attention for its wide applications in crowd management, intelligent surveillance and metropolis security. Counting the number of people manually is both time and labor consuming while machine learning methods provide an effective way for crowd counting. However, considering the factors such as occlusion, perspective and irregular arrangement, the automated counting of people is still a challenging task.

There existed a number of counting approaches in the past years [1, 2, 3]. Li et al. [1] estimated the number of people in surveillance scenes with people gathering and waiting through head-shoulder detection. Chen et al. [2] took the changing crowd conditions presented in different local regions into consideration, thus discovering the inherent relations among different features of spatial locations for crowd counting. Idrees et al. [3] utilized multiple sources such as low confidence head detections, repetition of texture elements to count people at patch level, and then enforced smoothness constraint on nearby patches, thus producing better estimates at image level.

Deep learning has been applied widely in computer vision, such as crowd segmentation [4] and scene analysis [5]. Some researchers made their contributions to the application of Convolutional Neural Network (CNN) in crowd counting and
achieved satisfactory results [6, 7, 8, 9]. Tota et al. [6] fused SIFT features and deep CNN features to implement crowd counting, thus getting better result than traditional methods that use hand crafted features. Wang et al. [7] improved the robustness of CNN-based network with some negative samples. Zhang et al. [7] plot perspective maps of all images before feeding them into the network and proposed a switchable training scheme based on density map and global count. Zhang et al. [9] designed a Multi-column CNN (MCNN) by utilizing filters with receptive fields (RFs) of different sizes. The MCNN is adaptive to variations in people/head size due to perspective or image resolution.

The size of receptive fields differs among different convolutional layers [10]. In this study, we modify AlexNet [11] with a combination of feature maps from different layers. Since the sizes of people’s heads differ greatly, we integrate feature maps of pool_3 and pool_5 before the regression step. The receptive fields of these two convolutional layers are different, thus can deal with different sizes of objects.

The advantages of the proposed framework are as follows:

1) The modified AlexNet is adaptive to heads with different sizes caused by perspective and image resolution. We utilize feature maps of different convolutional layers, and the multi-scale receptive fields can cope with different sizes of image blocks.

2) The network is initialized with well-trained parameters. We utilize parts of AlexNet and initialize it with the well-trained parameters.

3) The modified AlexNet is robust to some specific backgrounds. We add some negative samples such as trees and buildings into training dataset to enhance the robustness of the structure.

FRAMEWORK OF MODIFIED ALEXNET

Fig. 1 displays some patches that contain people’s heads. The areas occupied by objects differ greatly, which range from 6 to 264 pixels. Traditional CNN only uses feature maps in the last layer for classification or regression, thus providing receptive fields that have only one size. Zhou et al. demonstrated that the receptive fields of different convolutional and pooling layers differ greatly [10]. In addition, the shallow layers of Alexnet mainly detect edge features or the combination of some edge features [12]. Convolutional 3 and the later layers can detect some semantic features such as heads and wheels, which are more useful for crowd counting. Considering the importance of semantic features, the sizes of RF and the computing cost, we finally select feature maps of pool 3 and pool 5 for merging.

![Figure 1. Some patches of the training dataset.](image-url)
Figure 2. Structure of the modified AlexNet.

Alexnet was trained on Imagenet, which is a dataset that contains more than 1.2 million images for classification. The well-trained parameters are used to initialize the modified AlexNet. Figure 2 illustrates the structure of modified AlexNet.

As Fig. 2 shows, the structure from conv_1 to pool_5 of Alexnet is used to extract feature maps firstly. We then introduce a pooling layer following conv_3. Feature maps of pool_3 and pool_5 are merged for further feature extraction. After two full connection layers, the modified AlexNet finally outputs the number of people.

The input of the modified AlexNet is defined as \( X = \{ x_1, x_2, \ldots, x_K \} \), where \( x_k \) represents the kth sample of the dataset. The label of the dataset is defined as \( C = \{ c_1, c_2, \ldots, c_K \} \).

\[
F_k^3 = \Psi_1(x_k | \theta_1),
\]

In Eq. (1), \( F_k^3 \) represents the feature maps of the kth sample and \( \theta_1 \) is the parameter of the structure from input to conv_3. \( \Psi_1 \) is the mapping function that contains 3 convolutional layers and 2 pooling layers. Similarly, the formulas that map the features from conv_3 to pool 3 and pool_5 are:

\[
F_k^{33} = \Psi_{21}(F_k^3 | \theta_{31}), F_k^{35} = \Psi_{22}(F_k^3 | \theta_{32}),
\]

Where \( \Psi_{21} \) is a pooling function with kernel size 2 and \( \Psi_{22} \) is a mapping function that contains 2 convolutional layers and a pooling layer.

Then feature maps of pool_3 and pool_5 are merged as \( F_k^3 \) for further extraction, as Eq. (3) shows:

\[
F_k = \Psi_3(F_k^3 | \theta_3),
\]

\( F_k \) is the final feature of the kth sample and \( \Psi_3 \) represents the function of the 2 fully connection layers.

The output of the modified AlexNet is the number of people within a patch. After the feature extraction process, we adopt a linear transformation to approximate the number of individuals in patch as Eq. (4) shows:

\[
\hat{c}_k = \Phi(F_k) = \theta_4^T F_k,
\]
In the training process of the modified AlexNet, Euclidean distance is selected as the loss function of the counting task, which can be formulated as:

$$\ell(x_i, \Theta) = \frac{1}{2K} \left\| \hat{c}_i - c_i \right\|_2^2,$$

(5)

Where $\hat{c}_i$ and $c_i$ represent the output and the label of the kth sample. L2 norm is widely used to restrict the magnitude of the parameters for that it could prevent overfitting and thus improving the generalization ability of deep model [13]. In this work, we also use L2 norm as the weight penalties while training the modified AlexNet. Eq. (6) is the objective function for updating the parameters of the modified AlexNet:

$$\Theta' = \arg \max_{\Theta} \left( \sum_{i=1}^{K} (\ell(x_i, \Theta) + \|\Theta\|^2) \right).$$

(6)

During the training process, the initial learning rate is defined as $3 \times 10^{-6}$, and will be reduced by 0.1 every 20000 iterations.

**EXPERIMENTS**

**Dataset**

We conduct the experiments on part A of Shanghaitech dataset, which is the largest dataset for dense crowd counting [9]. The dataset contains 300 samples for training and 182 samples for testing. The number of people in one image ranges from 66 to 2256. People’s heads are annotated with dots (one person with one dot). This dataset contains a wide range of scenes, such as square, street, stadium and natural sceneries. Fig. 3 (a) displays some representative images of this dataset. Inspires by [7], we use some negative images that only contain backgrounds such as sceneries and architectures to train the modified AlexNet. We searched for different scenes from Google image search engine and finally selected 36 images as negative samples. Fig. 3 (b) displays some representative images of the positive and negative samples.

![Figure 3](image_url)

Figure 3. Representative images of (a) positive and (b) negative samples.
Experimental Results

The dataset is separated into training and testing samples, we use the training and the negative samples to train the modified AlexNet framework. MAE and RMSE are utilized as the evaluation criteria.

\[
MAE = \frac{1}{M} \sum_{m=1}^{M} |\hat{c}^m - c^m|, \tag{7}
\]

\[
RMSE = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{c}^m - c^m)^2}. \tag{8}
\]

Where \( M \) the number of testing is images in the dataset, \( c^m \) and \( \hat{c}^m \) is the ground-truth and estimation value of the mth testing sample respectively. Roughly speaking, MAE indicates the accuracy of the estimates and RMSE indicates the robustness of the estimates.

We compare our method with some deep learning methods and ridge regression (RR) method based on Local Binary Pattern (LBP) features. Table 1 displays the comparison results. As Table 1 shows, deep learning methods out-perform handcrafted features. Zhang et al. [8] tagged the perspective maps of the images manually before feeding images into CNN-based framework and used similar scene retrieval in testing process. But the tagging process was a hard task and their work was more effective for low density crowd. Zhang et al. [9] used multi-column CNN to cope with different sizes of heads. They labelled the density maps according to the distance between a person and the K nearest neighbors, but the labelling method is not adaptive to crowds with different densities. We also conduct experiment on the same dataset with Alexnet [11]. From Table 1 we can see that feature maps with different sizes of receptive fields can increase the estimation accuracy effectively.

To compare the estimation results in detail, we divide the testing samples into 10 groups according to the crowd counts in an increasing order. Fig. 4 displays the ground-truth and estimation results of these 10 groups. As Fig. 4 shows, our method outperforms other methods on crowds with both low and high densities, especially for group 4-8 (with an average of 86-506 people). For extremely dense crowds (group 9 and 10, with an average of 1603 and 2256 people), the modified AlexNet underperforms some of other deep learning based methods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>303.2</td>
<td>181.8</td>
<td>110.2</td>
<td>114.3</td>
<td>101.7</td>
</tr>
<tr>
<td>RMSE</td>
<td>371.1</td>
<td>277.7</td>
<td>173.2</td>
<td>173.6</td>
<td>155.4</td>
</tr>
</tbody>
</table>

Table 1. Comparison with other methods.
To display the counting results more intuitively, we give some specific patches and the estimations in Fig. 5. Fig. 5 (a) is a patch with pure background, and modified AlexNet can distinguish background from crowd more effectively. From Fig. 5 (b) we can see that attributing to the multi-scale receptive fields, modified AlexNet is more adaptive to the size of people’s heads. Fig. 5 (c) and (d) contain both crowd and backgrounds, and there exist occlusion. In this case, modified AlexNet still performs better than Alexnet. The crowd in Fig. 5 (e) is dense, and modified AlexNet can generate more accurate estimation than Alexnet. For extremely dense crowd, modified AlexNet performs similar to Alexnet, as Fig. 5 (f) shows.

CONCLUSION

In this paper, we modified AlexNet for dense crowd counting. The sizes of people’s heads in images vary greatly. To deal with this problem, we merged feature maps which have different sizes of receptive fields that could provide semantic features from different layers. Parts of Alexnet were used as feature detector, thus providing the network with better initialized parameters. In addition, we searched for some negative samples to enhance the robustness of the method. The experimental results on Shanghaitech dataset demonstrate that the proposed method outperforms traditional method which uses handcrafted features. Compared with other CNN-based
method, the proposed method gets the state-of-the-art performance by merging feature maps with different sizes of receptive fields.

ACKNOWLEDGEMENTS

The authors would like to thank the support from China Natural Science Foundation (No.61273277, No.61673244).

Corresponding author: Faliang Chang, flchang@sdu.edu.cn, School of Control Science and Engineering, Shandong University, 17923 Jingshi Road Jinan, P. R. China 250061.

REFERENCES