Pedestrian Counting via Deep Convolutional Neural Networks In Crowded Scene

Jingwei Li[1, a],Jianxin Song [1, b]

[1] Nangjing University of Posts and Telecommunications, NangJing 210000, China.

[a]lijw92@163.com , [b]iamsongjx@139.com

Keywords: pedestrian, counting, crowd, CNN, EIN.

Abstract. Currently pedestrian counts mainly faces two major problems in the crowd scene: expression of pedestrian's features and perspective shade. To address this problem, we propose a deep convolutional neural network (CNN) and Ensemble Inference Network (EIN) for crowd pedestrian recognizing and counting. First, we propose a Perspective-Correct Interpolation model for extract more features in crowded scenes. Then, we design a convolution layer which contains preventing occlusion layer in the neural network. Finally to achieve lower error rates, our CNN-based method introduces Ensemble Inference Network to the training and classification processes. Experimental results show that the error rate of the proposed method is 0.28 and possess accurate recognition in pedestrian comparable performance to state-of-the-art methods in crowded scenes.

1. Introduction

The current crowed counting has two major types: (1) Quantitative estimate the approximate number of crowded pedestrian and the average absolute error or counting accuracy is used as statistical accuracy metrics. (2) The population is divided into different density levels estimated, and by the different density levels estimates the pedestrian counting.

(1) Quantitative estimate

This method is represented by Davies, quantitative estimate pedestrian counting in a crowded scene, and counting accuracy is used as statistical accurate metrics. Counting by global regression ignores spatial information of pedestrians. Lempitskyet al. [1] introduced an object counting method through pixel-level object density map regression. Following this work, Fiaschi et al. [2] used random forest to regress the object density and improve training efficiency. Besides considering spatial information, another advantage of density regression based methods is that they are able to estimate object counts in any region of an image. Barandiaran J [3] proposed population count based on real-time multiple virtual lines, loading camera up the head, set a series of virtual line of statistical regional for avoiding occlusion occurs. We calculate the number of pedestrian from the region and out of the region, the rate of correct results 90%. However, these methods possess low accuracy statistic pedestrian as the occlusion is more popular in the crowded scene.

(2) Density level estimation

These methods represented by Marana, the population is divided into "lower", "low", "medium", "high", "higher" five density levels or other represented density levels through the size of the crowded density. M. Rodriguez [4] use density map estimation to improve the head detection results. These methods are scene specific and not applicable to crowded counting. Wu X [5] realized classification and texture features study estimates based on population density, then the image is divided into a small unit with the perspective projection model calibration, and SVM classification and density anomaly detection by texture feature vectors extracted from each unit.

Ouyang W [11] proposed a deep convolutional neural network (CNN) for shade scene counting, and it is trained alternatively with two related learning objectives, crowd density and crowd count. However, this method need density distribution previously obtained. The CNN employs a Rectified Linear Unit as an activation function, and Hiroshi Fukui uses Dropout to obtain a generalization [6] and the Dropout randomly removes a fixed ratio of units and Ensemble Inference Network (EIN) to the training and classification processes.

The key technology is the crowded pedestrian counting: (1) Finding the right expression feature of pedestrian in a crowded scene, and analyze using a suitable learning algorithms or regression algorithm. The common features are such as shape, size, edge, texture, motion trajectory. (2) There are so much perspective and occlusion (or overlapping) phenomenon in the crowded scene [7]. (3) The error rate of current methods is still high, and pedestrian are many similar characteristics in the crowd. So we could think of the thought of statistics for prediction classification results, such as: means, median, max [8].

First, we propose a Perspective-Correct Interpolation model for extract more features in crowded scenes. Then, we design a convolution layer which contains preventing occlusion layer in the neural network. Finally To achieve lower error rates, our method introduces Random Dropout and Ensemble Inference Network (EIN) to the training and classification processes after CNN layer.

2. Related work

Pedestrian detection is an important problem in computer vision, and there are the following three main directions for solving those problems.

2.1 Vision Correction

In the surveillance video, it is obvious that farther scene away from the camera there are smaller pixels in the image station, and the closer, there are more pixels. Pedestrians have great differences between the three-dimensional image with the same size two-dimensional, pedestrians closer the distance between the camera occupies a higher proportion of image pixels [9]. Space projective deformity is inevitable result created from three-dimensional to two-dimensional, and is a very important reason impacting of population monitoring system. So, removing projective affect could improve the accuracy of the system. Based on consideration of this reason, we proposed different scale parameter to describe different scale characteristics of population characteristics when analyzed population characteristics. Now there are two available correction projective deformity algorithm: perspective correction parameter and linear interpolation. They have advantages and disadvantages [10]. The former has a high degree of precision, but contains massive computation and complex. The latter has a lower Accuracy, but contains fewer calculation and better real-time performance.

2.2 CNN

CNN is a deep learning model, and could automatically learn and extract features from the data. Its generalization ability was significantly better than the traditional method has been successfully applied to the pattern classification, object detection, recognition and other fields [11, 12]. The convolutional neural network is a multi-layered supervised learning network with an input layer, a hidden layer (layer including convolution and down sampling layer) and the output layer. It acquires optimize the network structure by error back propagation (BP) algorithm to solve the unknown parameters. Its network structure is shown below:

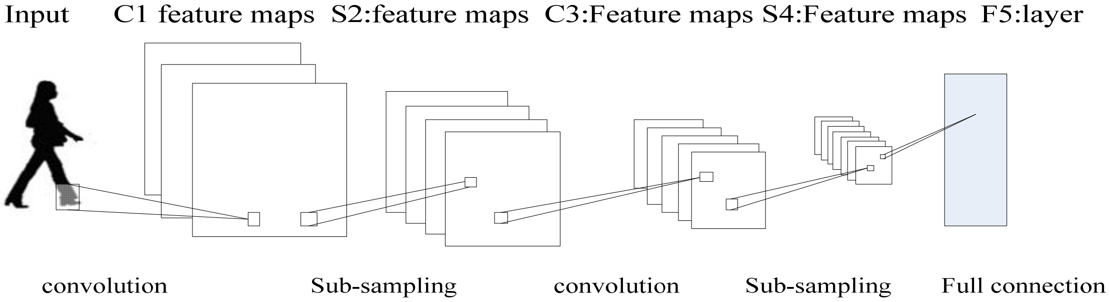

Figure 1 CNN structure

In the convolution layer, each neuron of feature map is connected to the previous layer receptive field, and the layer extracts local features through convolution operation. Convolution layer possessed a plurality of feature maps, and each map feature extracting a feature. The same feature mapping neurons share a same convolution kernel while extracts features. Different characteristic maps possess different weights, whereby different features could be extracted. The layer constantly adjusted the weighting parameters, so that feature extraction toward in favor of direction of classification in the process of training. the convolution layer is created as:

$$X_j^1 = f(\sum_{i\in M_j} X_i^{l-1} \times K_{ij}^l + b_j^l) \tag{1}$$

Where l is the number of layers, K is on behalf of the convolution kernel, M_j is an input layer receptive field, b is a behalf of bias.

In Down Sampling layer, number of input characterized mappings are unchanged after pooling layer, its size becomes the original 1/n (we suppose size of pooling was n). Pooling main role is to reduce the resolution of characteristic, decrease the feature dimensions, and increase network displacement, scaling, distortion robustness. Down Sampling Layer formula as show Eq(2):

$$X_j^l = f(\beta_j^l down(x_j^{l-1}) + b_j^l) \tag{2}$$

Where down (.) is the pooling function, β is the weighting factor.

2.3 Statistical Prediction.

In a crowded scene, mostly pedestrians are in blocked state, and we have found certain statistical properties. Pedestrian detecting based on convolution neural network possess a certain error. So we add a whole layer connection after convolutional neural network for more accurately detecting pedestrians. In training, the whole connection layer unit weights set to 0 with probability 1/2 after the final convolution visibility layer network random selection. We calculate statistical characteristics as the mean, median, maximum, etc. for determine whether a pedestrian.

3. Method

In crowded scenes, we firstly use interpolation for correcting position and decreasing the effects of camera. Then the convolutional neural network train and identify pedestrians. We designed a model to identify the crowded occlusion scene in the network convolution layer. Finally, adding a whole connection layer improve accuracy rate. Indicated as follows:

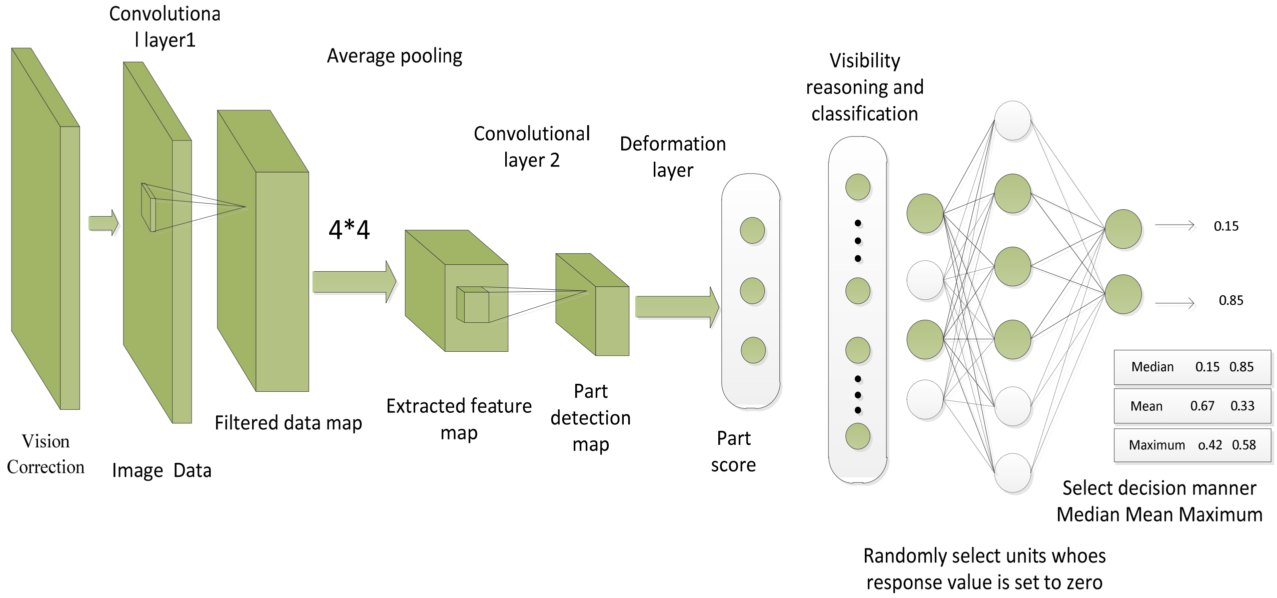

Figure 2 counting process of the crowd

3.1 Interpolation correction.

The traditional perspective correction algorithms give smaller weight for shorter distance from the camera and major right far away from the camera pixels. For crowded scenes, there are massive occlusion, especially multiply pedestrians overlapped, and less extractable feature, so we set greater weight for distribution overlap region and smaller weights for the target pixel thinning region. Corner density is represented by ρ , $\rho = N / S$, where S represents the area of $d \times d$, d is the pixels Euclidean distance, N is the number of intersection points within the neighborhood S , ω represents weight right center pixel S . d generally take the diameter of the non-overlapping pedestrian occupied area. The procedure is as follows:

(1) Motion segmentation obtains foreground binary image;

(2) Harris Corner detection on the foreground binary image, obtaining an outlook Harris Corner picture;

(3) From left to right and from top to bottom, scan foreground Harris Corner picture, obtaining the total number of corners for each pixel $d \times d$ neighborhood. According to the relation between the W and p, the each foreground pixel is assigned a weight to correct occlusion, obtaining a foreground pixel block diagram correction weight.

3.2 Occlusion model

Observed the occlusion pedestrian in a crowded scene, we propose the occlusion detection model that reference Joint CNN model [11]. The first layer sets 28 * 84 size picture as a first convolution layer input in convolutional neural network, A first convolution layer possesses 64 filter, each neighborhood enter data convolves with 9 * 9*3 filter parameters and outputs 64 feature maps. Then we obtained an average pooling of the 64 filtered data maps by using 4 × 4 boxcar filter. In the second convolution layer, we obtain 20 part detection maps through convolving the feature maps with 20 part filters of different sizes. Then part scores are obtained from the 20 part detection maps using a deformation handling layer. Finally the visibility reasoning of 20 parts is used for estimating the label y.

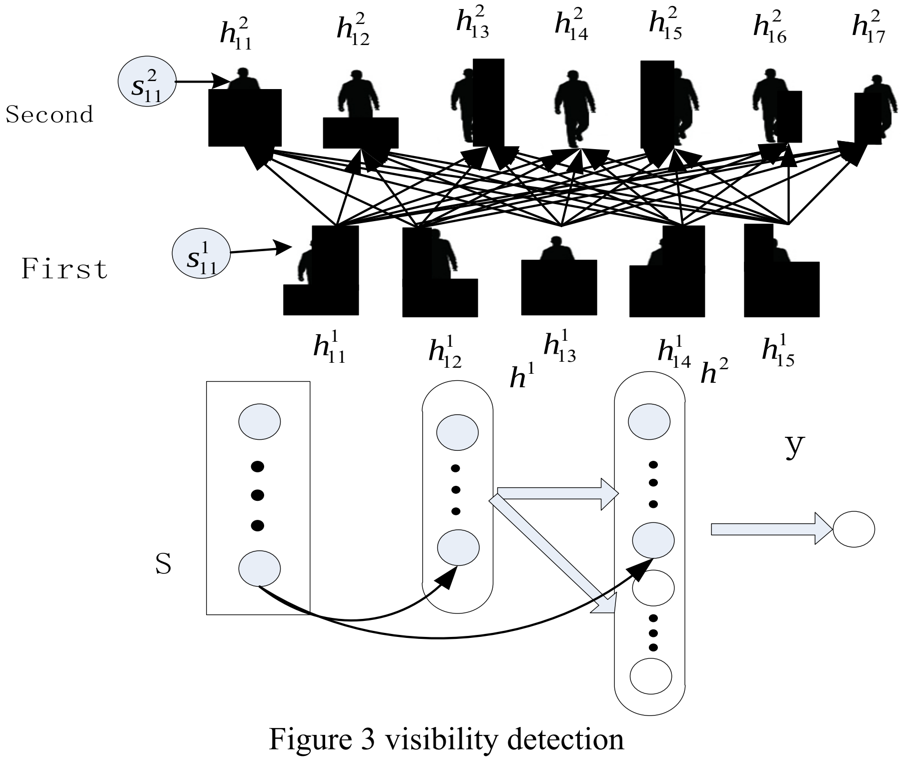

Figure 3 visibility detection

First, we estimate for component visibility of the first stage, and infer component visibility of second stage according to the visibility of the first stage [13].

H represents the visibility of the corresponding parts, y represents the final discriminant label, y is obtained from the following:

$$\widetilde{h}_j^l = \sigma(c_j^l + g_j^l s_j^l)$$

$$\widetilde{h}_j^{l+1} = \sigma(\widetilde{h}^{lT} W_{*,j}^l + c_j^{l+1} + g_j^{l+1} s_j^{l+1}), l = 1,2$$

$$\widetilde{y} = \sigma(\widetilde{h}^{3T} w^{cls} + b) \tag{3}$$

Wherein $\sigma(t)=(1+\exp(-t))^{-1}$ is Sigmoid function, g_j^l is the Weight parameters of s_j^l ,s_j^l is the bias, $\widetilde{h}_j^i$ represents the jth component ith stage visibility, W^l represents the correlation coefficient h^l and h^{l+1} ,$W_{*,j}^l$ is a jth column collection of element w^l , w^{cls} is a linear classifier of hidden unit $\widetilde{h}^2$,b is bias, g_j^l 、 c_j^l 、 W^l 、 w^{cls} and b are obtained by studying.

3.3 EIN

There are a lot of similar occlusion pedestrian characteristics in a crowded scene crowd. So we randomly select some of the units after the second components visibility and weights set up 0. Finally, we compare pedestrian detection outcome including median, average, maximum value for prediction [14]. The procedure is as follows:

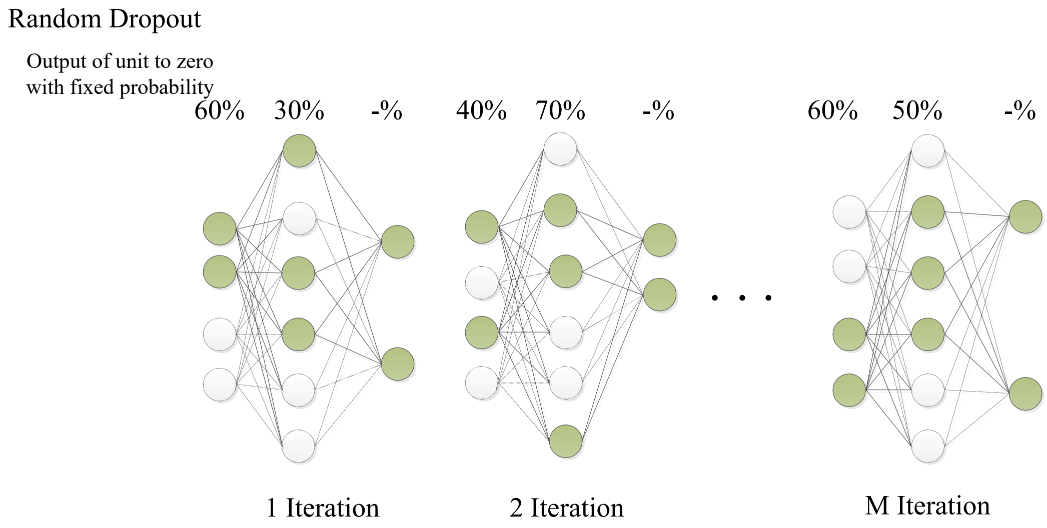

Figure 4 ensemble inference network

Feature map. The input visible unit I convolve with filter V , generating activation function, as follows:

$$h=\phi(V^T I+b) \tag{4}$$

b is bias, In K feature map of each unit, we select the maximum value as the active output. as follows:

$$h_i^{'}=\max_{k\in[1,K]} h_{ik} \tag{5}$$

Selecting the maximum value for each certain region, and improving the robustness of the change in small shape by resetting the feature Map size. As follows:

$$h_i^{''}=\max_{p\in P_i} h_p^{'} \tag{6}$$

Construct network. In the whole connected layer and the layer classification, the response of the randomly selected unit is set to zero. As show (7):

$$h_j^{(l)}=\phi(W^{(l)}x+b^{(l)})\cdot m_j^{(l)} \tag{7}$$

Visibility feature map obtained in step 2 is defined as x. $W^{(l)}$ is the weight, $b^{(l)}$ is bias, m controls the response value $h_j^{(l)}$ and is set to 0 or 1. Randomly selects N units is set to 0. N network classification probability calculated by the following:

$$O_{nc}=\frac{\exp(W_j^{(L)}h_j^{(L)}+b_j^{(L)})}{\sum_{c=0}^{C}(W_c^{(L)}h_c^{(L)}+b_c^{(L)})} \tag{8}$$

Classification: Each unit will calculate O_{nc} save as S_c collection, and selecting the median value as output result.

4. Experiments

In order to verify the accuracy of this algorithm in the crowd scene, the experiment is conducted in LIVE video and Caltech database. The databases mainly include roads and streets, living in the scene, and marked individual pedestrian and crowd. We also compare the proposed method with HOG [15], JCNN [11], Random Dropout EIN [8] for Caltech Pedestrian Dataset. In this paper, experimental tool is MATLAB2013b, CPU2.30GHz, memory 6.00GB.

Table 1 structures of network

Numble	property	Caltech Dataset	LIVE video
Input	size	84*28*3	84*28*3
conv1	Weight filter	9*9*3	9*9*3
pooling2	Map Size	19*5*64	19*5*64
Conv2	Weight filter	4*4	4*4
Pooling2	Map Size	20	20
Deform	Part	20	20
EIN	# of Fully connected	1000	1000
	# of Fully connected	500	500
	# of Fully connected	100	100
Output	Softmax	2	2

The structures of CNN+EIN are shown in table 1. The Caltech Pedestrian Dataset has 4000 positive samples and 20000 negative samples for training. The CNN include Forward Propagation (FP) and Backward Propagation (BP). FP firstly, process image as the input of the input layer, then action with each neuron of layer, and finally sent to the output layer, obtain an output. The BP adjusts layer weight matrix based on minimizing error. EIN has three Fully connected layers after component visibility layer , output has two units.

The TABLE 2 is performance comparison and experiment result with Caltech database, In this table, we compare our test set accuracy against best reported results in the literature. We note that for this database, our method actions superior performance compared favorably against state-of-the-art algorithms in a crowd scene. Wherein, FPPI is False Positive Per Image.

Table 2 comparison of Miss Rate

Algorithm	Miss Rate (FPPI=0.1)
HOG	53.42%
JCNN	39.92%
RandomDropout+EI	37.77%
N	28.54%
Proposed method	

5. Summary

In this paper, we have investigated the ability for CNN joint EIN to learn features from video frames. We propose a Perspective-Correct Interpolation model for extract more feature and design a convolution layer which contain preventing occlusion model layer, and introduces Random Dropout and Ensemble Inference Network (EIN) to the training and classification processes. Experimental results show, our method possess error rate of 0.2854 lower than the current algorithm. With the accuracy rate, the time consumption is also increasing. So, Future research directions are how to reduce the time consumption for real-time counting.

References

[1] Lempitsky, V., & Zisserman, A. Learning to count objects in images. In Advances in Neural Information Processing Systems (2010).(pp. 1324-1332).

[2] Fiaschi, L., Köthe, U., Nair, R., & Hamprecht, F. A. Learning to count with regression forest and structured labels. In Pattern Recognition (ICPR), 2012 21st International Conference on (2012, November). (pp. 2685-2688). IEEE.

[3] Barandiaran, J., Murguia, B., & Boto, F. Real-time people counting using multiple lines. In 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services (2008, May).(pp. 159-162). IEEE.

[4] Rodriguez, M., Laptev, I., Sivic, J., & Audibert, J. Y. Density-aware person detection and tracking in crowds. In 2011 International Conference on Computer Vision(2011, November). (pp. 2423-2430). IEEE.

[5] Wu, X., Liang, G., Lee, K. K., & Xu, Y. Crowd density estimation using texture analysis and learning. In 2006 IEEE international conference on robotics and biomimetics (2006, December). (pp. 214-219). IEEE.

[6] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580. (2012).

[7] Chen, Q., Jiang, W., Zhao, Y., & Zhao, Z. Part-based deep network for pedestrian detection in surveillance videos. In 2015 Visual Communications and Image Processing (VCIP) (2015, December). (pp. 1-4). IEEE.

[8] Fukui, H., Yamashita, T., Yamauchi, Y., Fujiyoshi, H., & Murase, H. Pedestrian detection based on deep convolutional neural network with ensemble inference network. In 2015 IEEE Intelligent Vehicles Symposium (IV) (2015, June).(pp. 223-228). IEEE.

[9] Wu, B., & Nevatia, R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (Vol. 1, 2005.pp. 90-97). IEEE.

[10] Kong, D., Gray, D., & Tao, H. Counting Pedestrians in Crowds Using Viewpoint Invariant Training. In BMVC.(2005, September).

[11] Ouyang, W., & Wang, X. Joint deep learning for pedestrian detection. In Proceedings of the IEEE International Conference on Computer Vision(2013). (pp. 2056-2063).

[12] Sermanet, P., Kavukcuoglu, K., Chintala, S., & LeCun, Y. Pedestrian detection with unsupervised multi-stage feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2013). (pp. 3626-3633).

[13] Tian, Y., Luo, P., Wang, X., & Tang, X. Deep learning strong parts for pedestrian detection. In Proceedings of the IEEE International Conference on Computer Vision (2015). (pp. 1904-1912).

[14]Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. IEEE, 2005.