An Effective Approach to Crowd Counting with CNN-based Statistical Features

Shunqiang Liu
Anhui University
School of Mathematical Sciences
Hefei, China

Sulan Zhai
Anhui University
School of Mathematical Sciences
Hefei, China

Chenglong Li
Anhui University
School of Computer Science and Technology
Hefei, China

Jin Tang
Anhui University
School of Computer Science and Technology
Hefei, China

***Abstract*—Recent works on crowd counting have achieved promising performance by employing the Convolutional Neurol Network (CNN) based features.These works usually design a deep network to detect pedestrian heads, and then count them. In this paper, we propose a novel approach to count pedestrians effectively based on the statistical CNN features. In particular, our approach only uses the first layer features of the CNN pre-trained offline on ImageNet, and thus obtains an efficient solution for crowd counting. Then, by analyzing the statistical properties of the first layer features, we observate the number of people fluctuates according to the value of the statistical features. Therefore, we employ these statistical features to train SVM, and can thus directly obtain the number of pedestrians. Experimental results on standard benchmark, UCSD, verify the effectiveness of the proposed approach.**

***Index Terms*—Support Vector Machine, Convolutional Neural Network, Crowd counting, statistical feature.**

I. Introduction

COunting crowd in videos draws a lot of attention because of its intense demands in video surveillance. Crowd counting is a challenging task due to severe occlusions, scene perspective distortions and diverse crowd distributions. Most works on crowd counting have solved the problem how to count them via the traditional methods. Then recent works on crowd counting have achieved promising performance by employing the Convolutional Neurol Network.

Many traditional works [1]–[3] focus on hand-crafted features or features selection, and employ these features to detect the objects. According to the number of objects, they count them. Because features are limited by local information, local segmentation and manual intervention, the detection accuracy of these previous works can also be promoted. Existing algorithms on crowd counting rely on hand-crafted features to calculate the number of crowd or population density. Existing regression methods can be divided into three categories: pixel-based analysis[4,5]; texture-based analysis[6, 7]; and object-level analysis [8], [9]. These methods [10], [11] rely on some global or local image features(mainly shapes,edges)to count them, and they are extracted by hand-crafting, selection features and manual intervention. However, recently many works have employed deep learning on crowd counting, and have achieved promising performance.

Deep learning models on crowd counting [12], [13]compute the number of people in the image by object detection. These works use the whole and online CNN net to detect objects, and then estimate the number of people or classify population density into certain types. Deep learning works on crowd counting achieve more promising performance than previous traditional works by employing the Convolutional Neurol Network (CNN) based features. The whole CNN net includes at least five convolutional layer, two fully layer and one classifier, which is a complex process, online and not end-to-end. So our work employs a simple and offline CNN net and achieves promising performance.

Thus,this paper present an effective and simply approach with CNN-based statistical features to solve the problem of crowd counting. In our approach, we extract the first layer features of the CNN pre-trained offline on ImageNet, and then employ the SVM to count them. The remainder of the paper is organized as follows: the second section reveiws related works of counting crowd. The third section introduces the proposed framework and its main components. More details of approach and implementation are also discussed. Extensive experimental evaluations and comparisons are presented.

Our approach employ the statistical CNN features and SVM to count them. Our works divid into three parts. First, we extract the first layer features on the pre-trained and offline ImageNet. Then, based on the first layer CNN features, and we employ the statistical first layer features(such as: mean, variance)as the input of the SVM. Finally, The SVM compute the number of people with the mean and variance first layer features, and detail process shows in Fig. 1.

II. Related work

In recent years, deep learning is very popular and widely applicated in the field of image and video. On crowd counting, many models applied the features of deep learning [14], [15]to count them. Many works introduced that deep learning is

978-1-5386-2524-8/17/$31.00 ©2017 IEEE

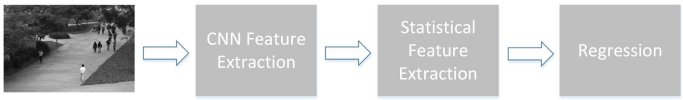

Fig. 1. This is the flow diagram of our training.

employed on various surveillance applications, such as person reidenti?cation [16], pedestrain detection [12], [13], tracking [17], crowd behavior analysis [18]and crowd segmentation [15]. Their works showed that they have achieved promising performance by employing the Convolutional Neurol Network (CNN) based features. However, existing deep learning algorithms employ all convolutional and fully layer of convolutional neural network to extract features,and use the features to detect object or to classify.And they need train the pre-trained CNN net with a large number of their data.

Traditional works on crowd counting are to creat the hand-crafting features,or tomake human intervention on feature selection. These works on crowd counting [1]–[4], [6], [7] employ hand-crafting features or features of huaman intervention. These models need human intervention to extract their features. Pixel based methods employs local features such edge information to count the number of people.Texture based methods depends on texture modelling through the analysis of image patches, include grey-level co-occurrence matrix, Fourier analysis and fractal dimension.Many works proposed to count the number of pedestrians by detection [19], [20]. But on crowd counting,these methods are limited by severe occlusions between people. A lot of methods [2], [5]predict the number of people by regression trained with low-level features. [14]introduced convolutional neural network to compute the number of pedestrians.Convolutional neural network has applied in many fields and achieves promising performance.

This paper provides an effective approach to crowd counting with CNN-based statistical features. Furthermore, The contribution of our approach is in three aspects. First, our framework only uses the first layer features of the CNN pre-trained offline on ImageNet, and thus obtains an efficient solution for crowd counting. Second, we extract the statistical CNN features as the input of classifier or regression. Third, this approach achieves to directly compute the number of people via the SVM, and does not using detection to obtain the number of pedestrian.

III. Our Approach

In this section, we introduce our approach on crowd counting, and detailedly overview the main components. This approach is mainly divided into two parts. we extract the CNN-based statistical features from the first layer features of the CNN pre-trained offline on ImageNet, and then predict the number of people in the image via the SVM with the statistical features. To introduce the structure of our approach at detail, the pipeline of our approach shows in Fig. 2.

Given an input image, this approach extract the first layer features of the CNN pre-trained offline on ImageNet, and then extract the statistical CNN features. We employ the SVM to count them with the statistical CNN features. In the following content, we will introduce the details of this approach.

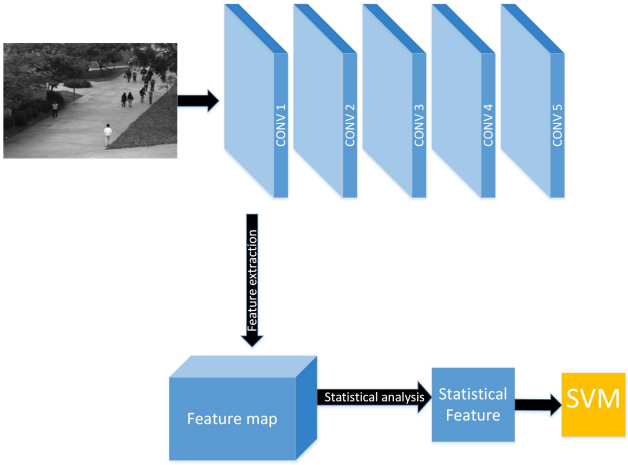

Fig. 2. the structure and process of our system.

A. Deep Convolution Features

In recent years, deep learning is employd in the field of detection and recognition.In the field of detection and recognition,deep learning has defeated many traditional methods. [14] shows the advantages of deep learning features and powerful function of the CNN net.We extract the first layer features of the CNN pre-trained offline on ImageNet(the structure of The convolution network shows in Fig. 3).

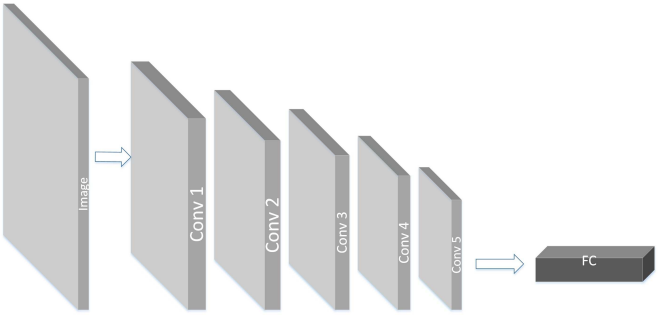

Fig. 3. The structure of traditional convolution network

In this approach, we extract the first layer features of the whole image on ImageNet, and ImageNet is pre-trained. On the ImageNet, there are five convolutional layer and two fully layer, which are feature extractor and feature selector. But we only extract the first layer features on the pre-trained ImageNet, and the first layer feature is a high dimension vector. The first layer features on the pre-trained ImageNet are these simple features(as: shape edge) or combination of these features. These first layer features can obtained by the first layer on the pre-trained ImageNet.

In this paper,our approach chooses the first layer features in the pre-trained ImageNet, and then extract the first layer

features in the pre-trained ImageNet. The vector shows that there is 96 feature extractor in the first layer, where every feature extractor extracts the 55×55 dimensions feature.Thus we obtain a $96 \times 55 \times 55$ vector from the first layer of the CNN pre-trained of fiine on ImageNet.

B. Statistical Ananlysis of Deep Convolutional Features

These paper [21], [22]show statistical analysis has employed on translation, and deep learning has used histogram(a kind of statistical features). So these paper proves efficiency of statistical features on traditional works. From these paper, we think we can apply the statistical deep learning features to count the number of pedestrian. Statistical features are widely used on the traditional methods and machine learning.But statistical features are not commonly used on deep learning. Thus in this paper, we use statistical deep learning features to count crowd in deep learning. In our approach, we extract the statistical first-layer features on the pre-trained ImageNet. We can obtain the high dimensional vectors.In this paper, we choose two statistical features, which are mean and variance. So we extract the mean and variance of the first layer features on the pre-trained ImageNet.

$$m = \frac{1}{3025}\sum_{i=1}^{55}\sum_{j=1}^{55} x_{ij} \qquad v = \frac{1}{3025}\sum_{i=1}^{55}\sum_{j=1}^{55} (x_{ij} - m)^2 \quad (1)$$

x_{ij} is 55×55, m is the mean of 55×55 vector, and v is the variance of the vector. Because the vector is a hige dimension, we reduce dimension of the vector, which is the statistical first layer features on the pre-trained ImageNet, and is a n dimensional vector. Thus we employ PCA(Principal Component Analysis) to reduce the high dimensional vectors.After reducing dimension,we build one vector of low dimension composed of the statistical CNN features. Our approach apply the SVM to compute the number of people with the low dimension vector.

C. Crowd Counting via SVM regression

In our approach, we employ the SVM(support vector machine) to count them. The SVM is the most common and effective linear and nonlinear regression and classification methods.Due to the nonlinear relationship between the number of people and the statistical first layer features, we employ the kernel function of SVM to compute the number, and the input of SVM is the statistical first layer features. In the training phase,the input of SVM is the vector, which mixes mean and variance first layer faeatures. we employ the mixed vector to train the SVM. The mixed vector is input of the SVM, and represents the mean and variance first layer features on the pre-trained offline ImageNet. Next section, we will introduce the experiment of our approach.

IV. Experimental Results

A. Datasets

In this paper, We employ the publicly available UCSD [23]crowd counting dataset to compare our results to previous work. We evaluate our approach on the UCSD datasets. This dataset contains 2000 images, and provides the ground truth with the number of people in every image. As is shown in Fig.4 from UCSD. And the details of the UCSD dataset is in table.1, which introduce the number of frames, resolution, density and total of number of pedestrain instances.

Fig. 4. Example frames from the datasets:Row is sampled respectively from the UCSD dataset

Data	N_p	R	D	T_p
UCSD	2000	238×158	11-46	49885

TABLE I
N_p=number of frames,R=Resolution,D=density(minimum and maximum number of people in the ROI), and T_p=total of number of pedestrain instances.

B. Evaluation Metrics

We employ two evaluation metrics, namely mean absolute error (MAE) and mean squared error (MSE) to evaluate our model.The two metrics can show the efficiency and accuracy of the model. For both metrics, the lower the values are, the better the experimental performance is. Thus we apply the two metrics to evaluate our model. The calculation formula of MAE and MSE shows in following.

$$MEA = \frac{1}{N}\sum_{i=1}^{N_f} |y_i - \widehat{y_i}| \qquad MSE = \frac{1}{N}\sum_{i=1}^{N_f} (y_i - \widehat{y_i})^2 \quad (2)$$

where N is the total number of test images.And y_i is the predictive value of my system in the image. $\widehat{y_i}$ is the true number of people in the image.The MEA reflect the accuracy of the approach.The MSE performs the stability of the framework. Then we count the number of people with the mean and variance first layer features, and we compare our approach with the other algorithms in UCSD dataset.

C. Comparison Results

Under the framework based on the regression, we try to compare the performance of different image representations with our approach. In order to prove promising performance of features parts on our proposed framework, we conduct a series of baselines:1 Holistic Feature (HF). The image representation is obtained by direct mean pooling over the entire dense attribute feature map. 2 SPP Feature (SPPF). The image representation is similar to HF except applying spatial pyramid on the entire feature map.3 LAF + VLAD (LFV). The image representation is gained by using the original VLAD method to encode our proposed locality-aware features LAF. The comparison results of the experiment are in the Table 2. Table 2 lists results by previous methods and our approach.

Method	MAE	MSE
MLR [9]	2.60	10.10
NCA-RR [2]	2.85	11.9
RFR [2]	2.85	8.47
VLAD [24]	2.41	9.12
HF	3.51	18.7
SPPF	3.47	17.46
LFV	3.37	18.14
Our approach	2.29	9.05

TABLE II
It can be seen that MAE/MSE achieves 2.29/9.05, which is a comparable result.

D. Component Analysis

In UCSD dataset, our approach employs the mean and variance first layer CNN features in pre-trained offline ImageNet to compute the number of people by SVM regression, and we employ the kernel function of SVM. Then experiment proves that the statistic first layer CNN features is useful and promising performance. Thus the result of the experiment shows that our approach has achieved promising performance by employing the statistic convolutional neurol network (CNN) based features. The advantages of our approach are that we use statistical first layer features in the pre-trained offline ImageNet, and can directly count them by SVM regression. However, our limitations are two parts. First, the performance improvement of our approach is not obvious. Second, the accuracy of our approach has declined in complex environment.We compare the statistic first layer CNN features with the first layer CNN features, and discover that the MAE of the first layer CNN based features to compute the number of people is 9.7. So we employ the statistic first layer CNN features to count crowd.

V. Conclusion

In this paper,we have presented a simple yet eddfctive approach for crowd counting, where our approach only uses the first layer features of the CNN pretrained offiine on ImageNet, and extract the statistical CNN features. Based on the statistical CNN features, we employed the SVM regression to count crowd. We extensively evaluated our approach on the UCSD dataset comparing with perious works. Experimental results demonstrated the advantages of our approach in accuracy and efficiency of our approach. In future work, we intend to incorporate segmentation techniques into our framework to further improve the accuracy of our method.

Acknowledgment

This study was funded by National Nature Science Foundation of China(61602002).

References

[1] T. Zhao, R. Nevatia, and B. Wu, Segmentation and tracking of multiple humans in crowded environments, IEEE transactions on pattern analysis and machine intelli- gence, vol. 30, no. 7, pp. 1198C1211, 2008.

[2] K. Chen, S. Gong, T. Xiang, and C. Change Loy, Cumulative attribute space for age and crowd density estimation, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2467C2474, 2013.

[3] V. B. Subburaman, A. Descamps, and C. Carincotte, Counting people in the crowd using a generic head detector, in IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, pp. 470C475, 2012.

[4] V. Lempitsky and A. Zisserman, Learning to count objects in images, in Advances in Neural Information Processing Systems, pp. 1324C1332, 2010.

[5] K. Chen, C. C. Loy, S. Gong, and T. Xiang, Feature mining for localised crowd counting., in BMVC, vol. 1, p. 3, 2012.

[6] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, Multi-source multi-scale counting in extremely dense crowd images, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2547C2554, 2013.

[7] W. Ma, L. Huang, and C. Liu, Crowd density analysis using co-occurrence texture features, in Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on, pp. 170C175, IEEE, 2010.

[8] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, Data-driven crowd analysis in videos, in 2011 International Conference on Computer Vision, pp. 1235C1242, IEEE, 2011.

[9] X.Wu, G. Liang, K. K. Lee, and Y. Xu, Crowd density estimation using texture anal- ysis and learning, in 2006 IEEE international conference on robotics and biomimet- ics, pp. 214C219, IEEE, 2006.

[10] K. Zu, F. Liu, and Z. Li, Counting pedestrian in crowded subway scene, in Im- age and Signal Processing, 2009. CISP09. 2nd International Congress on, pp. 1C4, IEEE, 2009.

[11] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, Privacy preserving crowd monitor- ing: Counting people without people models or tracking, in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1C7, IEEE, 2008.

[12] X. Zeng, W. Ouyang, and X. Wang, Multi-stage contextual deep learning for pedes- trian detection, in Proceedings of the IEEE International Conference on Computer Vision, pp. 121C128, 2013.

[13] X. Zeng, W. Ouyang, M. Wang, and X. Wang, Deep learning of scene-speci?c classi?er for pedestrian detection, in European Conference on Computer Vision, p- p. 472C487, Springer, 2014.

[14] C. Zhang, H. Li, X. Wang, and X. Yang, Cross-scene crowd counting via deep convolutional neural networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833C841, 2015.
[15] K. Kang and X. Wang, Fully convolutional neural networks for crowd segmenta- tion, arXiv preprint arXiv:1411.4464, 2014.
[16] W. Li, R. Zhao, T. Xiao, and X. Wang, Deepreid: Deep ?lter pairing neural network for person re-identi?cation, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152C159, 2014.
[17] N.Wang and D.-Y. Yeung, Learning a deep compact image representation for visual tracking, in Advances in neural information processing systems, pp. 809C817, 2013.
[18] J. Shao, K. Kang, C. C. Loy, and X. Wang, Deeply learned attributes for crowded scene understanding, in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4657C4666, IEEE, 2015.
[19] M. Wang and X. Wang, Automatic adaptation of a generic pedestrian detector to a speci?c traf?c scene, in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 3401C3408, IEEE, 2011.
[20] Z. Lin and L. S. Davis, Shape-based human detection and segmentation via hierar- chical part-template matching, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 4, pp. 604C618, 2010.
[21] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, Pcanet: A simple deep learning baseline for image classi?cation?, IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5017C5032, 2015.
[22] K. Cho, B. Van Merri enboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using rnn encoder-decoder for sta- tistical machine translation, arXiv preprint arXiv:1406.1078, 2014.
[23] A. B. Chan and N. Vasconcelos, Counting people with low-level features and bayesian regression, IEEE Transactions on Image Processing, vol. 21, no. 4, p- p. 2160C2177, 2012.
[24] B. Sheng, C. Shen, G. Lin, J. Li,W. Yang, and C. Sun, Crowd counting via weighted vlad on dense attribute feature maps, arXiv preprint arXiv:1604.08660, 2016.