An Improved Convolutional Neural Network on CrowdDensity Estimation

Shao-Yun PAN [1,a], Jie GUO[1] and Zheng HUANG [1]

[1]*School of Information Security, Shanghai Jiao Tong university, China*

Abstract. In this paper, a new method is proposed for crowd density estimation. An improved convolutional neural network is combined with traditional texture feature. The data calculated by the convolutional layer can be treated as a new kind of features.So more useful information of images can be extracted by different features.In the meantime, the size of image has little effect on the result of convolutional neural network. Experimental results indicate that our scheme has adequate performance to allow for its use in real world applications.

1 Introduction

Stampede may happen while people are too overcrowding, which always results in casualties. In 2016, for example, the New Year's day in Shanghai, there are too many people gathered in the bund. It results in casualties due to no effective evacuation. In order to prevent such incidents, the crowd density detection in video detection system is important. We always expect the system to inform us when people's density reaches a certain range we gave. In this way, people can be evacuated in time, avoiding an accident.

In the past decades, many efforts have been made in the respect. It is proposed a method based on detection individuals [1]. But the severe crowd occlusion makes the approach not up to our expectations [2]. In general, texture feature and a classifier are used to estimate the crowd density. It introduced the grey level co-occurrence matrix(GLCM) of an image [3, 4]. Then the support vector machine (SVM)is used to analyse crowd density [4]. The feature is effect when the crowd is very dense. But when the light changed seriously, the accuracy is not high [2]. A new feature is introduced named Local Binary Patterns(LBP) feature [5, 6]. It can have little change when the light condition changes thus it has advantages over GLCM. A new approach is discussed [7]. It is based on optical flow and hierarchical clustering. However, the method is not very good when the pedestrians are long period static.

In recent years, deep learning is very popular. It has been more and more widely applied. The paper proposes a method based on the convolutional neural network(CNN) to analyse the crowd density [8]. However, the network is too low that it is not always good in all areas. In this paper, we combined LBP texture feature and convolutional neural network. The features learned by convolutional layer tend to be partial. It can be more global with deeper layers. So we add one more convolutional layer and fully connection layer which can get more deep features. The LBP features extract other information which enhances the resolution of images. It can improve the overall recognition of images, controlling the error in a smaller range.

The rest of this paper is organized as follows: in Section 2, we will present our model briefly. And then we will discuss the basic concepts of LBP texture feature and CNN used in our model. In Section3, we discuss the result of our experiment. A brief summary showed in section 4.

2 Model

In this section, we propose an approach which is how to combine Local Binary Patterns (LBP) and convolutional neural network in our model. It can improve the accuracy in the experiment. We then discuss basic knowledge of LBP and convolutional neural network used in our experiment.

The overview of our method is shown in Fig. 1. In this paper, we design the model that has three convolutional layers and pooling layers. Each convolutional layer follows a pooling layer in order to decrease parameters' numbers. It can reduce computation complexity. C1 contains three filters and we design 32 feature maps to output in first stage. In the second stage, 32 feature maps change to 64 that each feature map generates two in the convolutional layer. The pooling layer just changes the size of maps. It has no effect for numbers. After the last stage, 90 feature maps obtain to output. In order to get better results, the convolutional neural network contains two full connections. The features of local binary patterns input the first full connection with feature maps together. The classifier is trained until 200 optimizations and selects parameters of the lowest error one.

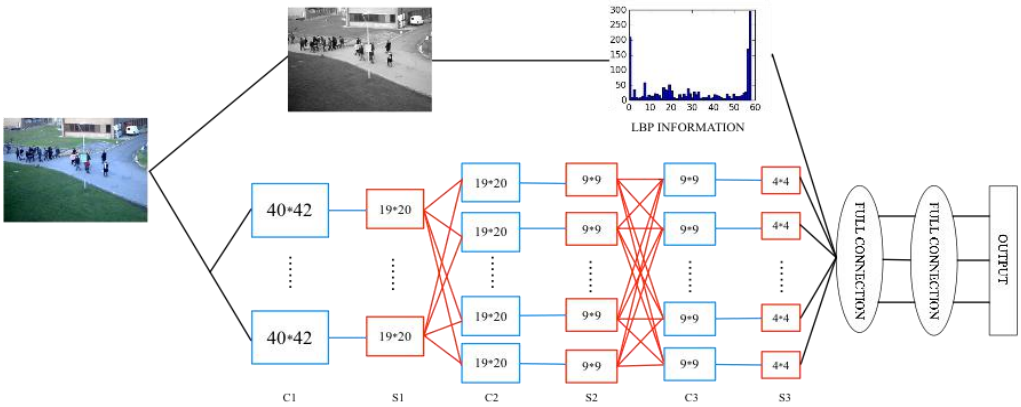

Figure 1. The structure of our model concludingthreeconvolutionallayers

2.1 Local Binary Patterns

Local Binary Patterns is a very common operator used to describe the image local texture features in the field of machine vision. LBP has some significant advantages, such as grey scale invariance and rotation invariance. It was firstly proposed by T. Ojala, M. Pietikäinen, and D. Harwood in 1994 [9]. LBP is very powerful in classification field when using features of images, so that it is widely used in image classification. Due to strong discrimination and simple to calculate, the operator of local binary patterns has been applied in many different scenarios.

On the side of texture analysis, image texture feature of a pixel refers to the relationship among this point and the surrounding pixels in most cases. In other words, it means the relationship of a point and its neighbourhoods. It will form different types of features after extracting feature from different angles. LBP constructs a measure of the relationship among a pixel and its surrounding pixels. LBP is a simple but very effective texture feature operator. Firstly, selecta pixel as centre. And then its neighbouring pixels will be compared to the central pixel. If the grey-level value of neighbouring pixel is larger, then the location is marked as 1, else marked as 0. Then every pixel will get a string of binary values. This is a process of converting the original image to LBP image. Finally, figure out its histogram.

LBP has a very good features. When the light of an image changes, the LBP feature extracted from an image changes little. So it has good robustness to illumination. This is the reason that it is more popular that other texture features

In this paper, we do not use the basic method because the binary modes are too many. Ojala proposed a uniform LBP [10]. With such improvement, even though the types of binary modes are greatly reduced, the information is not lost. Patternsare reduced as P*(P-1)+2 from the origin, where P represents the neighbourhood points. Figure 2 shows the procedure of extracting the feature. It can reduce the effect of high frequency noise, for dimensions of the feature vector are decreased.

2.2 Convolutional Neural Network

Convolutional neural network is a kind of artificial neural networks. It has become a hot topic in speech analytics and image recognition field. The weight sharing makes it more similar to the biological neural network. It reduces the complexity of network models and numbers of weight values [11]. This performance will be more obvious when dealing with multi-dimensional images. Images can input directly to the network, to avoid the complex feature extraction and data reconstruction in traditional recognition algorithm. Convolutional network is specially designed as a multi-layer perceptron to identify two-dimensional shape.

CNNs as a deep learning framework is proposed in order to minimize data in the preprocessing stage. Local receptive field is the input data in the lowest level of CNN structure[12]. Then the information is in turn transmitted to different layers. Each layer will obtain the most significant features through a digital filter. Local receptive field of an image can access to the most basic features by neurons and processing units.

Thereare two important parts in convolutional neural network. One is named alternating convolutional layer, and another one is pooling layer [13]. Generally, C layer is a layer of feature extraction. The input of each neuron is connected to the local receptive field of preceding layer, and extracting the local feature [14]. Once the local features are extracted, the positional relationship among it and other characteristics are also subsequently finalized. Computing convolution of theimage is actually a filtering process. Eq. (1) shows the process of convolutions.

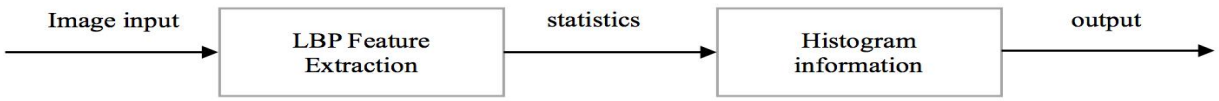

Figure 2. The procedure of LBP feature extraction

$$Y=sigmoïde(\sum x*w+b) \qquad (1)$$

Each feature map is calculated by a convolution kernel, and plus a bias. Finally, it calculates the sigmoid function. Eq. (2) defines the sigmoid function.

$$S(x)=1/(1+e^{-x}) \quad (2)$$

S layer is a feature mapping layer. Each layer is composed of multiple feature maps. Each of them is a flat, and weights of all neurons in the plane are equal. The layer is designed to reduce the feature map. Generally, to sum every four neighborhood pixels. The sum is weighted by w, and plus a bias b, and then activated by a function.

Due to sharing weights of all neurons in a same plane, the number of free parameters and the complexity of network parameters selection are reduced. In Convolutional neural network, each feature extraction layer is followed by a calculation layer used for local average and secondary extraction. It makes the network have a higher tolerance for distortion when identifying the input samples.

3 Experiment

In our paper, the Pets2009 dataset is used for experiment [15]. Table 1 shows the number of each class including training data and testing data.

Table 1. The number of images of each class.

Number	Very low	Low	Medium	High	Very high
training	120	141	137	97	217
testing	40	71	83	60	120

The dataset has three different crowd scenarios [16]. We use dataset S1 for density estimation. The dataset is not available directly. Firstly, according to the number of people in each image, images are divided into five categories. The range of number of people of each class refer to [8].

Table 2. Result of PETS_2009 by cascade optimized CNN and CNN combined LBP.

result	Very low	Low	Medium	High	Very high
CNN	100	88	87	78	90
LBP+CNN	100	91	92	88	92

Table 2 shows the result of our experiment. Our method getsbetter performance comparing with the method in Literature [8], especially for the low, medium and high class. The very low and very high class is almost the same. The main reason may be our approach takes advantage of both convolutional neural network and traditional features. Both of them have own advantage alone. Each of them play a role, forming complementarily in our model.

Table 3. Result of PETS_2009 by LBP and common CNN.

Result (%)	Very low	Low	Medium	High	Very high
LBP1	95	83	69	72	94
LBP2	85	79	56	28	96
CNN1	91	88	90	83	97
CNN2	87	87	92	81	95

Table 3 shows the result of experiment with images of different scales. LBP1 represents the result of images in large size, and LBP2 is for the small images. Others also are the case. The image size has a great impact on our result to the LBP feature. Almost the accuracy of all categories have decreased, especially for the kind of high. To the CNN method, the performance has a little change. Of course, we use three convolutional layers similarly, since the result is better than two.

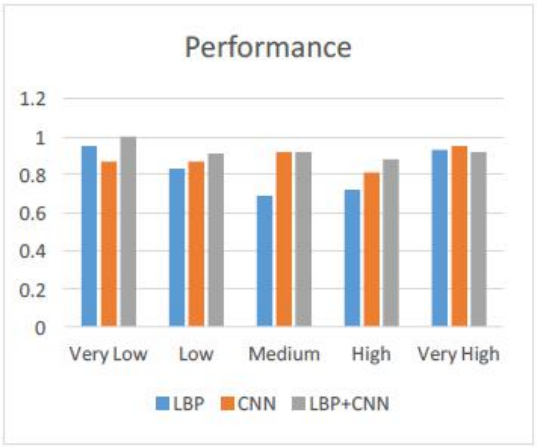

Figure3. Crowd density estimation performance of different methods Crowd density estimation performance of different methods.

Itis clearly shown in Fig.3 that our method has the better result. The performance is always stable. The LBP feature does work as well as the CNN. But when the size of images is very small, the feature of LBP is very bad. In other words, it depends on high quality images.The model is very useful when the crowd density is high. This information is important in reality.

4 Conclusion

Video surveillance has become an integral part in people's daily life. It's more and more widely used, because it can significantly relieve labor. The crowd density estimation which is an important component of abnormal behavior monitoring, can effectively prevent the loss caused by the crowd out of control. For the estimation of the crowd density problem, the most common method is texture analysis. The first step is to extract texture feature of the

image, and then select some appropriate classifier for training. Nowadays, deep learning has been greatly developed and applied to the crowd density estimation. It has been proved that the deep learning algorithm is more effective than the ordinary methods.

There are different approaches to estimate crowd density, but all of the methods have its own problems. The paper introduces some LBP information and convolutional neural network knowledge. Inspired by the idea of boosting algorithm, we present a new approach based on LBP and CNN in the paper.The resultcan be better when the two are combined, since the model extracts more features from images. These two kinds of information have little relationship. It gets stronger since combining two different features.

References

1. Rittscher Jens, Tu Peter H., Krahnstoever Nils, Simultaneous estimation of segmentation and shape, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **2**, 486-493(2005)
2. Saleh S.A.M, Suandi S.A, Ibrahim H., Recent survey on crowd density estimation and counting for visual surveillance[J]. Engineering Applications of Artificial Intelligence, **41**,103-114(2015)
3. Wu Xinyu,Liang Guoyuan, Lee K.K, Xu Yangsheng, Crowd density estimation using texture analysis and learning, 2006 IEEE International Conference on Robotics and Biomimetics, ROBIO, 214-219 (2006)
4. Wang B, Bao H, Yang S, et al. Crowd Density Estimation Based on Texture Feature Extraction[J]. Journal of Multimedia, **8(4)**: 331-337 (2013)
5. Wang Zhe, Liu Hong, Qian Yueliang, Xu Tao, Crowd density estimation based on local binary pattern co-occurrence matrix, Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops,ICMEW, 372-377 (2012)
6. Yang H, Su H, Zheng S, et al. The large-scale crowd density estimation based on sparse spatiotemporal local binary pattern[C]//2011 IEEE International Conference on Multimedia and Expo. IEEE, 1-6 (2011)
7. Rao Aravinda S, GubbiJayavardhana, MarusicSlaven, Stanley Paul, PalaniswamiMarimuthu, Crowddensity estimation based on optical flow and hierarchicalclustering, Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics, ICACCI, 494-499 (2013)
8. Fu Min, Xu Pei, Li Xudong, Liu Qihe, Ye Mao, Zhu Ce, Fast crowd density estimation with convolutional neural networks, Engineering Applications of Artificial Intelligence, **43**, 81-88, August 1 (2015)
9. Ojala T., Pietikainen M., Harwood D., Performance evaluation of texture measures with classification based on Kullback discrimination of distributions[C]//Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on. IEEE, **1**: 582-585 (1994)
10. OjalaTimo, PietikäinenMatti, MäenpääTopi, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**, n 7, 971-987, July (2002)
11. MrazováIveta, Kukacka, Marek, Hybrid convolutional neural networks, IEEE International Conference on Industrial Informatics (INDIN), 469-474 (2008)
12. LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision[C]//ISCAS, 253-256 (2010)
13. Sermanet Pierre, ChintalaSoumith, Lecun Yann, Convolutional Neural Networks Applied to House Numbers Digit Classification, Proceedings - International Conference on Pattern Recognition, 3288-3291 (2012)
14. Scherer D, Müller A, Behnke S. Evaluation of pooling operations in convolutional architectures for object recognition[C]//International Conference on Artificial Neural Networks, 92-101 (2010)
15. http://ftp.pets.rdg.ac.uk/pub/PETS2009/Crowd_PETS09_dataset/a_data/a.html
16. Saleh Sami Abdulla Mohsen, SuandiShahrelAzmin, Ibrahim Haidi, Recent survey on crowd density estimation and counting for visual surveillance, Engineering Applications of Artificial Intelligence, **41**, 103-114, May 1 (2015)