Pattern Recognition 48 (2015) 3016–3024

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Large scale crowd analysis based on convolutional neural network

Lijun Cao, Xu Zhang, Weiqiang Ren, Kaiqi Huang

Center for Research on Intelligent Perception and Computing (CRIPC) National Laboratory of Pattern Recognition (NLPR) Institute of Automation, Chinese Academy of Sciences (CASIA), China

ARTICLE INFO

Article history:
Received 8 October 2014
Received in revised form
6 March 2015
Accepted 2 April 2015
Available online 15 April 2015

Keywords:
Crowd analysis
Flow counting
CNN
Large scale data

ABSTRACT

Nowadays crowd surveillance is an active area of research. Crowd surveillance is always affected by various conditions, such as different scenes, weather, or density of crowd, which restricts the real application. This paper proposes a convolutional neural network (CNN) based method to monitor the number of crowd flow, such as the number of entering or leaving people in high density crowd. It uses an indirect strategy of combining classification CNN with regression CNN, which is more robust than the direct way. A large enough database is built with lots of real videos of public gates, and plenty of experiments show that the proposed method performs well under various weather conditions no matter either in daytime or at night.

1. Introduction

People always gather in high density at public places, such as city streets, subway stations or malls. Individual-level analysis loses its effectiveness in this situation. As a result, more and more researchers today in video surveillance pay their attentions on crowd-level analysis, especially on crowd counting [1]. Estimating the number of people in a surveillance scene is significant for many applications, such as market analysis or public security. However, it is still a challenging task because of the heavy occlusion, variant illumination and weather.

Plenty of methods have been proposed to deal with the problem of counting people. Most of the previous methods focus on solving the problem of region of interesting (ROI) counting. There are two ways for researchers to deal with the ROI counting: method by human detection or feature regression. The first method is usually incapable of working in crowed scenes or bad weather. For example, Lan et al. [2] proposed a background subtraction which maps the global shape feature to various configurations of humans directly. The algorithm is speeded up by MCMC and the experiments in the real scenes show its efficiency. However, it cannot handle the large crowded scene. Paul et al. [3] built a human detector, which is trained on adaboost and utilize both motion and appearance information to detect a walking person. The contribution of his work is that the method also works well on low-resolution images under bad weather, like snowy and rainy, but its effectiveness in real scene setting needs to be verified. Bo and Ram [4] proposed a part detector based on edgelet features, the responses of which are combined to form a joint likelihood model that includes cases of possibly inner-occluded humans. The experimental results on both previous datasets and new datasets, on which earlier methods cannot work well, show its outperformance compared with other human detection systems. For the feature regression methods, such as features edges, wavelet coefficients or a feature combination by some machine learning methods, they have a strong requirement for large amount of data. And the extracted features are heavily affected by camera perspective. Another disadvantage of ROI counting is that most of them cannot get the number of people with a different flow direction, and it limits its application in real scenes. Shengfuu et al. [5] combined the wavelet templates and head contour detectors to get the number of people in a crowded scene. However, the method is limited to some specific situations where the contours of head are clear enough. Chan et al. [6] presented a privacy-preserving system to estimate the size of regions in homogeneous crowds where pedestrians walking in different directions. He extracted features from connected regions and segments the image to get the number of people with Gaussian Process regression. Its performance is well in training data, but the generalization ability is quite poor.

Besides ROI counting, some methods focus on solving the problem of line of interesting (LOI) counting. Current state-of-the-art LOI counting approaches [7,8] are based on extracting and counting crowd blobs of the temporal slice of the video. Zheng and Chan [7] and Yang et al. [8] use the LHI dataset, which includes

E-mail addresses: lijun.cao@nlpr.ia.ac.cn (L. Cao), xu.zhang@nlpr.ia.ac.cn (X. Zhang), wqren@nlpr.ia.ac.cn (W. Ren), kqhuang@nlpr.ia.ac.cn (K. Huang).

http://dx.doi.org/10.1016/j.patcog.2015.04.001

videos with different views, to show that their methods can work well no matter the scene changing. However, counting large blobs containing many people is not accurate due to severe occlusion. Especially in real scene video surveillance, the handcraft feature can hardly handle the variation of weather, scene or illumination. As image features play an important role in either ROI or LOI methods, the recently booming CNN (Convolutional Neural Network), which is well-known by its ability of image feature representation, could offer an obvious progress in crowd counting.

CNN has enjoyed a great success in large scale visual problems, such as object detection and classification [9–13], image segmentation [10,14], face recognition [15,16] and others related problems [17–19]. As the development of large image dataset, such as ImageNet [20], and large scale distributed clusters [21], CNN becomes a panacea in computer vision field, a number of attempts have been made to improve the original architecture. For instance, Pierre et al. [9] introduced a novel deep learning approach to localize object by learning and predicting object boundaries. He used a single shared network to easily integrate recognition, localization and detection together and won localization task in the ImageNet Large Scale Vision Recognition Challenge 2013 (ILSVRC2013). Ross et al. [10] proposed a detection algorithm called Regions with CNN features (R-CNN). The detection algorithm improves the mean average precision by more than 30% compared with the previous best result on VOC2012. The 3D CNN also plays one of the main roles in computer vision task. Shuiwang et al. [22,23] proposed a framework of 3D CNN for human action recognition. This method extracts features from both spatial and temporal dimensions and has a competitive performance on the TRECVID dataset and KTH dataset. But 3D CNN needs much more time and memory in computing.

Taking advantages of CNN in image feature representation, we propose a crowd flow estimating method using LOI with few parameters, and this method can distinguish the number of entering and exiting people even in crowding scenes in various real environments. By obtaining the temporal slices of original images and optical flow images, we get redundant data containing entering and exiting people information, which is then inputted into a classification CNN and a regression CNN to compute the crowd flow types and statistics of crowds respectively. With all these outputs of CNNs, the method gives out the specific number of entering and exiting people in real time.

The contribution of this paper lies in these aspects:

1. We propose a novel method to estimate the number of entering and exiting crowd flow with three CNN models. Compared with traditional methods, it shows robustness under various real scenes and hardware settings. Furthermore, the approach takes an indirect strategy of combining classification CNN (for crowd flow types) with regression CNN (for crowd flow quality), instead of directly exploiting a regression CNN model to obtain the number of entering and leaving people. This framework performs much better than the direct approaches in lots of experiments.

2. We build a database with about 140 thousand temporal slice images, which are captured under 7 real scenes and ranging from several to dozens of people. The database is growing as the new applications are deployed.

3. We tested our method under all weather conditions of 7 scenes, which are caught from various unrestrained gates of Public Parks. These experiments show a promising prospect of the method even under the condition of crowded people flow or mal-weathers.

The rest of the paper is organized as follows. In Section 2, the details of the method including model designing are introduced. The dataset is presented in Section 3. In Section 4, experiments are conducted to display the effectiveness of the proposed method.

2. Architecture of the algorithm

As shown in Fig. 1, there are four modules in our crowd flow estimating system. Temporal slice module extracts slice images from original images (all the original images are processed with mosaic for the privacy protection) and optical flow images. Two regression CNN modules are trained on these temporal slices to predict the overall number of people and the ratio of entering to leaving people respectively. In order to improve the estimation performance under all conditions, we propose to adopt a classification CNN module to classify input optical flow images into four different types for further processing. By fusing the output of these three CNN models, we can obtain robust and accurate estimation of both the number of entering and leaving people.

2.1. Temporal slice module

Given an input video sequence, a temporal slice is formed by sampling the LOI over the original images or the optical flow images. The LOI is a fixed-width line, as shown in Fig. 1, and each column in the slice image corresponds to the LOI at a given frame. The temporal slice of original image is made up of the RGB pixel value of LOI, which has three dimensions and ranges from 0 to 255. The temporal slice of optical flow image is made up of the

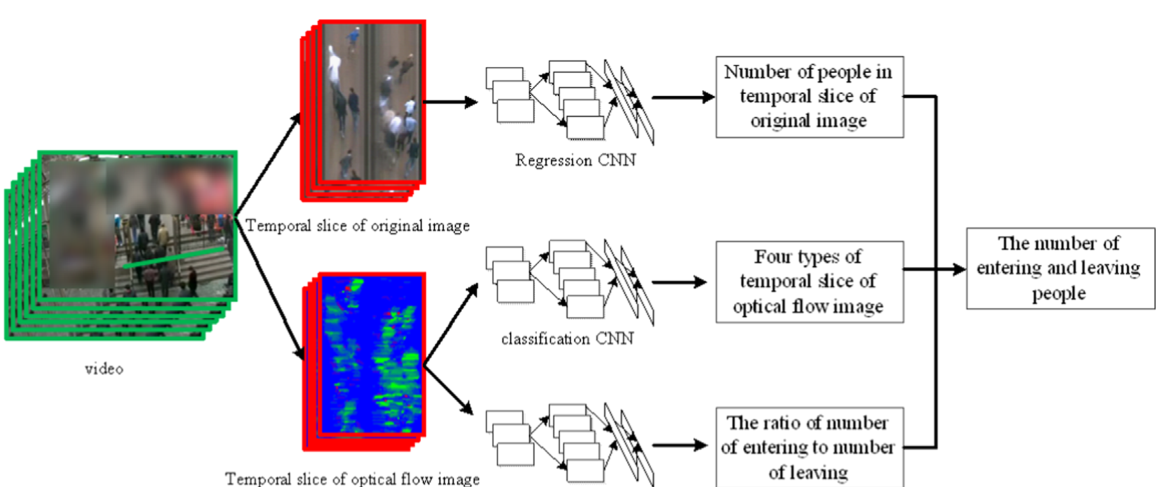

Fig. 1. Framework of the proposed algorithm.