Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2022.030205
ISSN 2790-0983 Vol. 3, Issue 2: 40-47

Automatic Head Count Based on Machine Learning in Intelligent Video Surveillance

Jiaqing Li*
Philippine Christian University, Philippine
ljq@sxu.edu.cn
**corresponding author*

Keywords: Video Surveillance, Machine Learning, Head Count, Target Tracking

Abstract: With the wide application of video surveillance system, visual information has become the key research element of modern security technology. Computer vision related technology can be applied to the field of intelligent surveillance, so that computers can process video. People can use computers to understand video surveillance, directly get the number of people in an area, or get the distribution of people. This paper first analyzes the existing two types of target detection algorithms, and chooses Fast R-CNN algorithm as the research object of this paper. This paper combines the research method of background modeling with the research method of deep convolutional neural network based on statistical learning to fuse all the calibration boxes of pedestrian detection results. A pedestrian count evaluation method is proposed, and the pedestrian count results are smoothed and fused.

1. Introduction

With the development of video surveillance technology, video surveillance can be seen everywhere. As a collection of pictures, video contains a large amount of information. Intelligent analysis based on the number of people data can help people improve efficiency [1]. In addition, with the growth of the population and the enhancement of people's security concept, the construction of national smart security and intelligent finance is continuously promoted, and the application of monitoring is increasingly widespread. The development of computer hardware makes it possible to process and analyze big data. The popularity of video surveillance makes monitoring more convenient. However, the current monitoring task of video surveillance is often manual operation, and analyzing and searching video often consumes a lot of manpower and material resources. Therefore, it is urgent to realize the intelligence and automation of monitoring, which has very important research value in real life [2-3]. Based on the video image analysis algorithm, accurate and effective statistics of the number of people in the video frame are carried

out, and various demand data reports are generated in the background of the system or in the client according to the change trend of the number of people at different times, which can help the staff to obtain instant information and make correct decisions in a limited time. The video head count algorithm can be widely applied to various public management places. According to the different crowd sizes, it can be divided into high-density head count and relatively low-density head count. The application scenarios of high-density head count include public transportation and passenger flow statistics, etc. The application scenarios of relatively low-density population statistics include classroom attendance, public areas, and exhibition booth passenger flow, etc. [4-5]. Specifically, for example, in the exhibition booth, when the system detects that the number of people has reached a certain value, it will automatically broadcast and explain. For example, the intelligent robot can judge the number of people in front of the vision, so as to provide corresponding services according to the number of people, etc., these tasks are based on accurate and fast statistics of the number of people [6].

In current methods based on object detection, the exact number of people is obtained by counting the results of human detection. Since the introduction of deep learning, this kind of method has made great progress. Methods based on object detection can be divided into One-Stage and Two-Stage methods according to the steps of whether to obtain the possible bounding box explicitly, and different solutions can be derived from different tasks based on them [7]. In the head count task, the head count task based on object detection can derive different research points according to human body detection based on the target detection milestone framework of stage 1 and stage 2, such as fusion of multi-layer features, direct detection on multiple convolutional layers, contextual information, local information, global information, etc. And combining these research points to improve one of the landmark detection framework, trying to improve in one or more aspects [8]. The Two-Stage method is based on a series of algorithms for candidate target regions, which continue the traditional process of target detection by first selecting candidate regions on images, and then doing classification and regression on them [9]. Some scholars used the method of convolutional network to participate in the competition and won the champion, which makes convolutional neural network applied to large-scale image classification task and emerge in the field of object detection. The classic work of the two stages is the R-CNN series algorithm, and its work is very continuous [10]. One-Stage is an end-to-end detection method represented by SSD (Single Shot MultiBox Detector) and YOLO (You Only Look Once). It takes target detection as a regression problem to eliminate time consumption in candidate areas to improve speed. At the same time, after a series of improvements, it is superior to the two-stage method of R-CNN series in the balance of accuracy and speed [11-12].

Real-time accurate video population statistics not only has its value in academic research and practical application, but also will further promote the field of computer vision research to the actual scene landing. This paper investigates the research status of video population statistics, explores the reasons for the main challenges in practical application, and puts forward corresponding targeted solutions to these challenges.

2. Video Population Statistics Based on Object Detection Algorithm

2.1. Object Detection Correlation Algorithm

Most of the backbone networks used in the existing target detection are convolutional neural networks, and the most common methods can be divided into two categories according to the different detection steps.

(1) Target detection based on candidate regions

The main implementation process of this detection algorithm is divided into two steps. The first

step is to obtain the candidate region, and the next step is to classify the object. There are many outstanding representative algorithms, such as R-CNN(Selective Search + CNN +SVM), SPP-Net, Fast R-CNN, and so on. Faster R-CNN et al., the average detection accuracy pairs of different algorithms are shown in Table 1 [13-14].

Table 1. Comparison of detection accuracy

Model	mAP(VOC2007)	mAp(COCO)
R-CNN	58.2	
SPP-Net	59.1	
Fast R-CNN	71.4	41.8

VOC dataset is a standard dataset, which is often used as a benchmark dataset for testing image classification. It contains 20 object classes, among which the scene of the picture is complex and changeable, which can well compare the advantages and disadvantages of the algorithm. COCO dataset mainly contains a large number of images, a total of 200,000 images, including 80 categories, and the test on this dataset is highly convincing. mAP represents the accuracy value of the algorithm detection, and the higher the value, the better the detection effect of the detection algorithm [15].

It can be clearly seen from Table 1 that, in the above two datasets, Faster R-CNN performs well in detection accuracy compared with other algorithms, but the detection takes a long time, mainly because the algorithm spends a long time on candidate box extraction, which eventually leads to slow detection speed. Compared with real-time target detection, Faster R-CNN has certain defects.

(2) Object detection based on regression

The difference between the regression-based method and the candidate region is that the idea of region is completely removed and the region suggestion network is not used. The regression and classification are completed in one network. Such difference makes the network less repetitive operations and finally performs well in terms of speed, which is suitable for fast detection scenarios. Classical methods mainly include YOLO and SSD [16]. YOLO algorithm mainly divides the detected image into S×S grids, which are then used to detect objects whose target center belongs to the grid. The basic network of SSD uses VGG16, which has strong extraction features. The network design integrates different scales and turns the last two fully connected layers into convolutional layers, and then forms the final network framework by adding four convolutional layers [17-18].

The comparison results of each algorithm are listed in Table 2. By comparing the results listed in the table, it can be seen that YOLO series algorithms perform well in data sets. Among them, YOLOv3 algorithm has a higher accuracy rate than other algorithms in COCO data sets in complex scenes, because it simplifies the algorithm complexity and has irreplaceable advantages in detection speed.

Table 2. Comparison of detection accuracy of regression algorithm

Model	mAP(VOC2007)	mAp(COCO)
YOLOv1	67.5	
SSD	75.2	50.2
YOLOv3		56.9

2.2. Fusion motion information neural network detection algorithm

The main structure of CNN is composed of many convolutional layers and pooling layers. In the convolutional layer, the neurons are not fully connected. In the same layer, mutual information sharing can be realized and the number of parameters in the training process can be reduced to a

large extent. The local perception feature is used mainly because researchers have found strong associations between regions that are close to each other in images, and weak associations between regions that are far apart. In the convolution operation, the step length is defined to represent the length of the distance, but the specific setting of the step length data has no fixed value and should be determined according to the actual application. In addition, the parameters of CNN can be shared, the parameters of neurons are the same, and the statistical characteristics of local information learned by the network can be applied to the structure of another network. At the same time, the information sharing of convolution kernel will also bring some problems, such as the incomplete and comprehensive extracted features, but multiple convolution kernels can also be used to solve this problem.

In this paper, a pedestrian video counting algorithm combining motion information and neural network static detection is proposed. Static image detection based on the depth of the convolution neural network frame, each frame of the pedestrian target detection at the same time with gaussian mixture background modeling algorithm to extract the image per frame prospect of pedestrian movement area information, then the two methods of extracting the pedestrian detection window do fusion processing after the window number as per frame video image to detect the line number, Finally, the number of rows in each frame of video image within 1 ~ 2 seconds is counted, and the number of pedestrians with the most occurrence is taken as the number of pedestrians detected at the end of the current frame. The main idea is: Based on the pedestrian region R detected by the deep neural network on the video image, the foreground moving target image region D and R extracted by Gaussian mixture background modeling are merged, and the merged region is used as the final detection result.

The algorithm proposed in this paper has certain requirements for the application environment, that is, it is suitable for the sparse target scene monitoring area. Because the hybrid Gaussian background algorithm and neural network are used to calibrate the whole human body, when the crowd in the monitoring scene is dense, there is a serious mutual occlusion effect of the crowd, which will have a great impact on the construction of the human body calibration frame. Secondly, under the sparse target scene monitoring area, the number of pedestrians in the video frame within 1 ~ 2 seconds can be considered to be basically stable. Under this assumption, the pedestrian count evaluation method is proposed.

Assume that xi is the real number of pedestrians in the current ith video image, si is the number of rows detected in the ith video image when the counting optimization algorithm is not adopted, and sl is the number of people detected in the ith video frame after the counting optimization algorithm is used to smooth the counting results. If the average accuracy of the statistics results of the number of video users with a certain number of video frames is calculated, it can be obtained as follows:

$$p = \frac{\sum_{i=1}^{N} s_i}{\sum_{i=1}^{N} x_i} \tag{1}$$

$$p' = \frac{\sum_{l=1}^{N} s_l}{\sum_{i=1}^{N} x_i} \tag{2}$$

Where, p is the accuracy rate obtained without counting optimization process, p 'is the accuracy rate obtained with counting optimization process, and Ⅳ is the number of video frames. It's not hard to see that p is less than p prime.

3. Video Population Statistics Experiment

3.1. The Experiment 1

The video data with moderate crowd density were selected for the experiment, and the reduced features were selected. The traditional Fast R-CNN neural network algorithm and the improved Fast R-CNN algorithm fused with motion information were used to analyze the recognition effect and time of each algorithm.

3.2. The Experiment 2

In order to verify the monitoring effect of the improved method used in this paper on the actual monitoring scene, two kinds of video data of the above simple scene and complex scene are selected for the experiment. Among them, the simple scene background is simple; Large human flow in complex scenes; Each group contains 2 video sequences to be tested. The improved Fast R-CNN algorithm fused motion information was used to calculate the recognition effect.

4. Analysis of Experimental Results

4.1. Results of Experiment 1

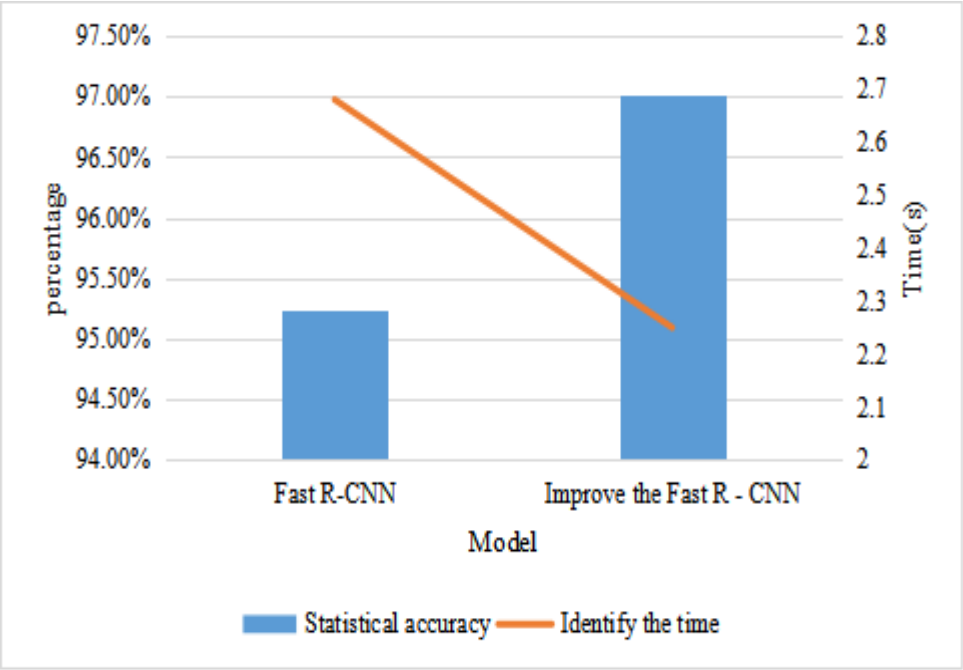

Figure 1. The improved algorithm compares the results

As can be seen from Figure 1, compared with the traditional Fast R-CNN algorithm, the improved Fast R-CNN algorithm with motion information fusion used in this paper has improved recognition time and statistical accuracy respectively.

4.2. Results of Experiment 2

The results of experiment 2 are shown in Table 3.

Table 3. Table of statistics results

	Actual number of people	Statistical number of people	Number of error checking	Number of residual
Simple scenario	26	24	1	1
	43	45	2	0
Complex scenario	32	29	2	1
	45	44	1	2

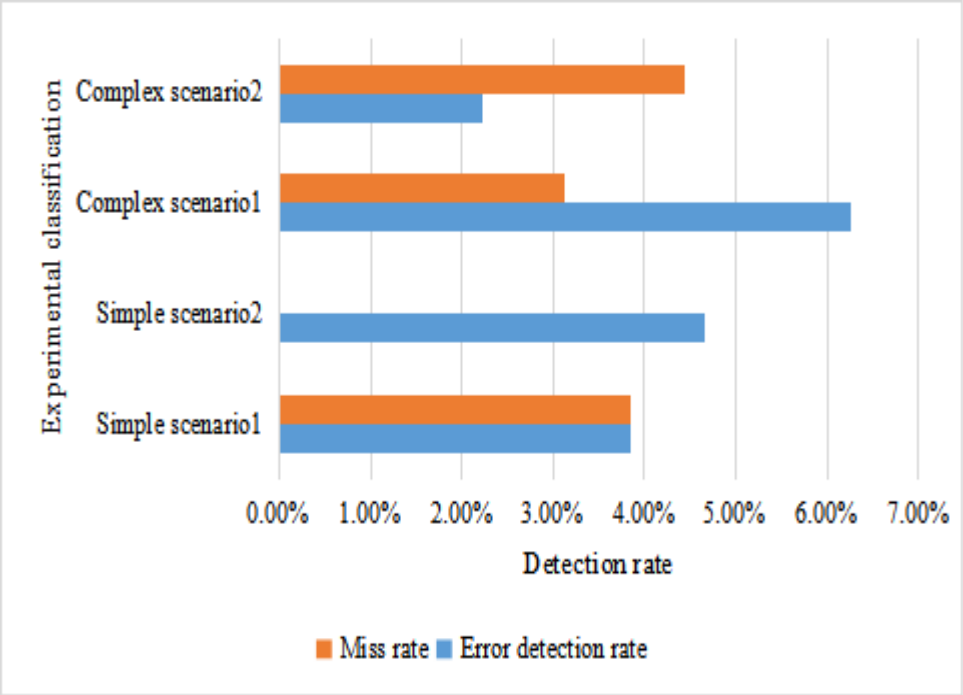

Figure 2. Statistics of false detection rate and missed detection rate

By analyzing FIG. 2, it can be seen that although the method used in this paper still has false detection and missed detection, they are all within an acceptable range. For complex scenes, it can also achieve high accuracy and has low false detection rate and missed detection rate.

5. Conclusion

With the rapid development and update of software and hardware technology, most of the video image processing technology can be realized, so many applications have entered people's life. The research and application of the number of people statistics based on intelligent video play an important role in real life. Compared with the traditional pedestrian detection and counting algorithm for extracting simple pedestrian features, this paper combines the research method of background modeling with the research method of deep convolutional neural network based on statistical learning to realize the video person counting system. The deep learning static image pedestrian detection results based on convolutional neural network are integrated with the foreground motion information extracted by the mixed Gaussian background modeling method, so as to improve the accuracy, robustness and universality of video pedestrian detection.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Hénaff O J, Bai Y, Charlton J A, et al. Primary visual cortex straightens natural video trajectories. Nature communications, 2021, 12(1): 1-12. https://doi.org/10.1038/s41467-021-25939-z

[2] Kowal M, Conroy E, Ramsbottom N, et al. Gaming your mental health: a narrative review on mitigating symptoms of depression and anxiety using commercial video games. JMIR Serious Games, 2021, 9(2): e26575. https://doi.org/10.2196/26575

[3] Austerschmidt K L, Stappert A, Heusel H, et al. Using a video presentation on variance and covariance in the teaching of statistics. Teaching Statistics, 2022, 44(1): 15-20. https://doi.org/10.1111/test.12292

[4] Valarmathi E R, Metilda M. Effectiveness of Video Assisted Teaching Programme on Knowledge about Polycystic Ovarian Syndrome among Adolescent Girls in a Selected School, Chennai. International Journal of Midwifery Nursing, 2022, 5(2): 30-34.

[5] Andoh J, Ebroff B D. Assessment of spontaneous eye blink rate in online livestream video game players. Adv Ophthalmol Vis Syst, 2021, 11(1): 11-14. https://doi.org/10.15406/aovs.2021.11.00401

[6] Agwi U C, Irhebhude M E, Ogwueleka F N. Video surveillance in examination monitoring. Security and Privacy, 2021, 4(2): e144. https://doi.org/10.1002/spy2.144

[7] Kumar A. Experience of video consultation during the COVID-19 pandemic in elderly population for Parkinson's disease and movement disorders. Postgraduate Medical Journal, 2021, 97(1144): 117-118. https://doi.org/10.1136/postgradmedj-2020-138846

[8] Bermúdez-Bejarano E, Bermúdez-Sánchez J A, Ruiz-Rey F J, et al. Analysis of psychic imbalance, caused by screening of a video of surgical extraction of a lower third molar in a sample of mental patients as compared to the general population. Journal of Clinical and Experimental Dentistry, 2022, 14(9): e726. https://doi.org/10.4317/jced.59861

[9] Permatasari A. The Effectiveness of Making English Video Related to Aquatic Resources Management Program Study in Improving The Students'Speaking Skill At Stip Muhammadiyah Sinjai. Agrominansia, 2018, 3(2): 163-171. https://doi.org/10.34003/281878

[10] Prabu I J, Anbumani J. A Study to Assess the effectiveness of Video Assisted Teaching Programme on the level of knowledge regarding Blood Donation among GNM 1st year students of AMT School, Jammu. International Journal of Advances in Nursing Management, 2020, 8(2): 127-132. https://doi.org/10.5958/2454-2652.2020.00030.X

[11] Mansoori M, Joshi K, Sharath S. A study to assess the effectiveness of video assisted teaching programme on knowledge regarding post-partum intra uterine contraceptive devices (PPIUCD) among antenatal mothers in selected rural areas at Udaipur, Rajasthan. IP Journal of Paediatrics and Nursing Science, 2021, 3(4): 118-121. https://doi.org/10.18231/j.ijpns.2020.022

[12] Woltsche R, Mullan L, Wynter K, et al. Preventing Patient Falls Overnight Using Video Monitoring: A Clinical Evaluation. International Journal of Environmental Research and Public Health, 2022, 19(21): 13735. https://doi.org/10.3390/ijerph192113735

[13] Hainsworth E, McGrowder E, McHugh J, et al. How can we recruit more men of African or African-Caribbean ancestry into our research? Co-creating a video to raise awareness of prostate cancer risk and the PROFILE study. Research Involvement and Engagement, 2022, 8(1): 1-7. https://doi.org/10.1186/s40900-022-00347-9

[14] Arnaez J, Vega-Del-Val C, Hortigüela M, et al. Usefulness of video recordings for validating neonatal encephalopathy exams: a population-based cohort study. Archives of Disease in Childhood-Fetal and Neonatal Edition, 2021, 106(5): 522-528. https://doi.org/10.1136/archdischild-2020-320791

[15] Byrne J V, Whitaker K L, Black G B. How doctors make themselves understood in primary care consultations: A mixed methods analysis of video data applying health literacy universal precautions. Plos one, 2021, 16(9): e0257312. https://doi.org/10.1371/journal.pone.0257312

[16] Slemmons K, Anyanwu K, Hames J, et al. The impact of video length on learning in a middle-level flipped science setting: Implications for diversity inclusion. Journal of Science Education and Technology, 2018, 27(5): 469-479. https://doi.org/10.1007/s10956-018-9736-2

[17] Alam R. The Use of Asynchronous Google Classroom with Video Teaching Tutorial on Improving Student Writing Achievement for Recount Text at Public Senior High School 6 Kendari. Journal of Applied Science, Engineering, Technology, and Education, 2022, 4(1): 80-87. https://doi.org/10.35877/454RI.asci884

[18] Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature, 2020, 580(7802): 252-256. https://doi.org/10.1038/s41586-020-2145-8