The Fifth International Conference on Railway Technology
RAILWAYS 2022
22–25 August 2022
Montpellier, France

Proceedings of the Fifth International Conference on
Railway Technology:
Research, Development and Maintenance
Edited by J. Pombo
Civil-Comp Conferences, Volume 1, Paper 17.4
Civil-Comp Press, Edinburgh, United Kingdom, 2022, doi: 10.4203/ccc.1.17.4

People counting in railway environment using computer vision

Sitou AFANOU[1]

[1]Rolling stocks engineering center, SNCF Voyageurs
Le Mans, France

Abstract

Ensuring passengers' safety and comfort is a daily comfort for railway operators. In order to correctly size their transport plan, it is necessary to know in real time the number of users present in the trains. With the advent of artificial intelligence, several techniques now make it possible to detect and track passengers effectively. The combination of these 2 techniques can then make it possible to carry out a relevant count. In this article, we compare 2 counting methods based on the one hand on the detection of faces, on the other hand on the segmentation of shapes and their reidentification. We will show the promising results obtained, as well as the impact of physical phenomena such as occultation on the precision of the counting performed.

Keywords: re-identification, detection, counting, tracking, computer vision, segmentation.

1 Introduction

The deployment of “Artificial Intelligence” in several domains has achieved big successes. Various emerging applications were thus been designed to become part of our daily lives. In the railway domain, AI can be used to improve the quality of services offered or to achieve things that were not possible before, especially using CCTV systems [1].

For instance, as part of the experimentation and industrialization of various “smart surveillance” systems , the rolling stock must embed completely autonomous systems,

which are used to analyze the flows and behavior of travelers on board: presence and density of travelers, number of people getting on and getting off at each station.

Various approaches have been proposed to tackle the problem of people counting in videos, log et al [2] broadly classified traditional counting methods into the following categories: detection base approaches, regression based approaches and density estimated based approaches. Recently CNN and attention based methods have shown great results in crowd counting and density estimation [3].

Among different fields of action, a counting system with infrared sensor has been used on each trainset, to ensure the security and the comfort of travelers. Indeed, the fact of being able to know the load rate for a given trip at a given time makes it possible to predict the number of trains required in advance to avoid overloading the train and to ensure passengers' comfort.

However, current systems are limited in terms of the reliability. To improve those results, we proposed two methods for people counting. Therefore, our contribution is twofold. First, it provides a high as well as improved performance system. Second, it brings real value added features by using the people re-identification.

Our paper is organized as follows. In section II, we will survey the different algorithms used for detection, and re-identification of passengers. Section III presents the results obtained using the algorithms described in section II, tested on two datasets, and the chosen algorithm tested in real time on a Z2N suburban train with numerous services and movements in stations. Section IV emphasizes on new and important aspects of the study and the conclusions drawn from them.

We propose two different approaches for passenger counting on board. The first one consists in counting passengers by detecting their faces, and the second one by segmenting and re-identifying them.

2 Methods

The use of these two approaches was to avoid the occlusion and collapse of bounding boxes by using others algorithms such as YOLO [4].

As well, redundancy is a problem that can distort our counting. In fact, each trainset is set up with two cameras, one at the entrance and the other at the exit, so at each time two frames are captured and analyzed, then passengers are counted. In case of segmentation, and to avoid counting a person in both frames twice, we need an algorithm that compare its mask extracted from a frame with all masks of the other one; hence the necessity of re-identification.

Counting with segmentation and re-identification:
Segmentation:

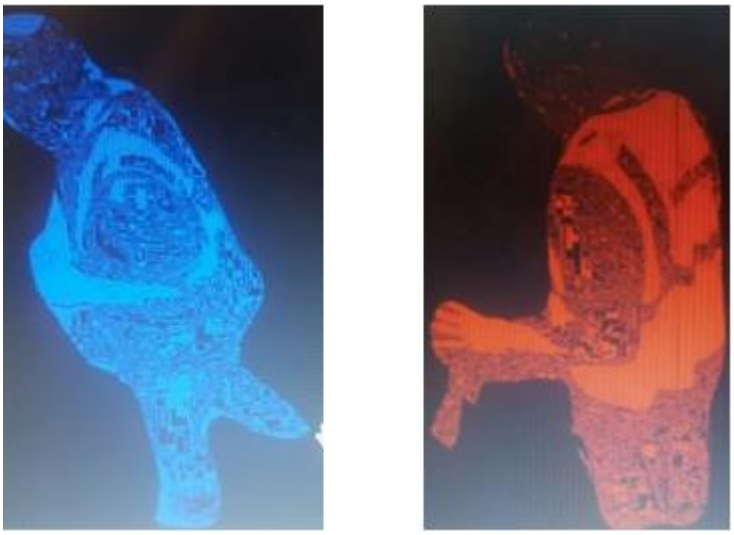

Figure 1: two persons segmented with MaskRCNN.

Re-identification:

Once all persons in the two frames were segmented, an algorithm of re-identification is applied to eliminate those that have been captured twice. We have experimented two existing algorithms named AlignedReid [6] and PersonReid [7], chosen for their good performances. The first one calculates the distance of two person image which is the sum of their local and global distance. Hence, if it is greater than 0.5, then the two persons are different, if not it is the same person. The second one enables the feature extractor to be aware of the interdependency of the matching items that can directly influence the computation of each other's representation.

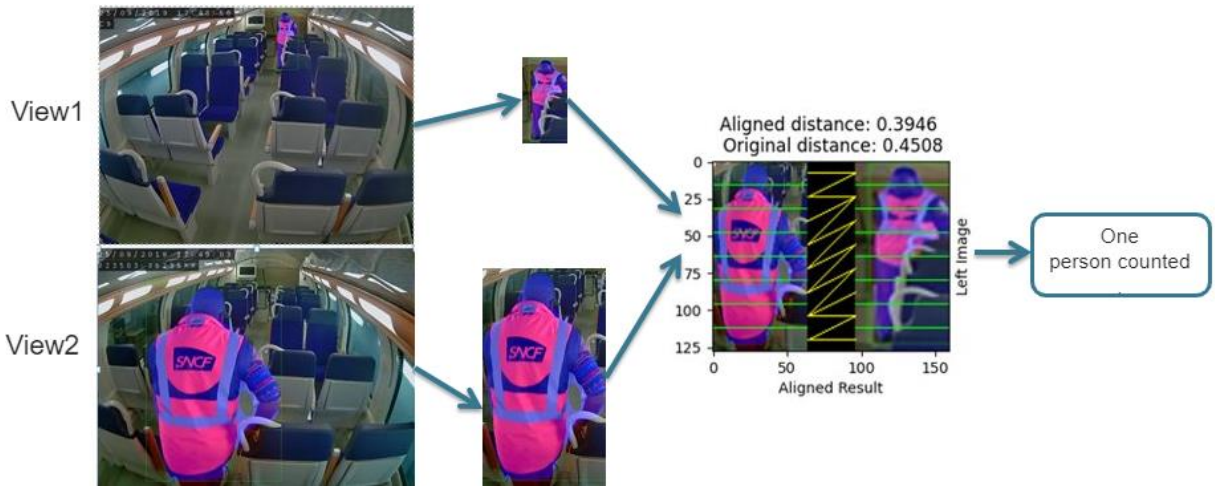

Figure 2: counting with segmentation and re-identification (AlignedReID+MaskRCNN)

2. Counting with face detection:

For this approach, it is assumed that a person detected in one frame is not detected in the second one, since his face will only be captured by the camera in front of him, while the second will capture his head. Consequently, the counting is realized by summing the number of faces detected in the two frames, see Fig 3.

We experimented with MTCNN [8], HaarCascad [9] and faceResNet_101 [10]. Though ResNet101 out performs the state of art since it deals with the problems of scale variation, image resolution and contextual reasoning. In fact, they train binary multichannel predictors to report object confidence for range of size, then they find large and smaller faces with a Coase image pyramid. Finally, ResNest101 is used for shared CNN.

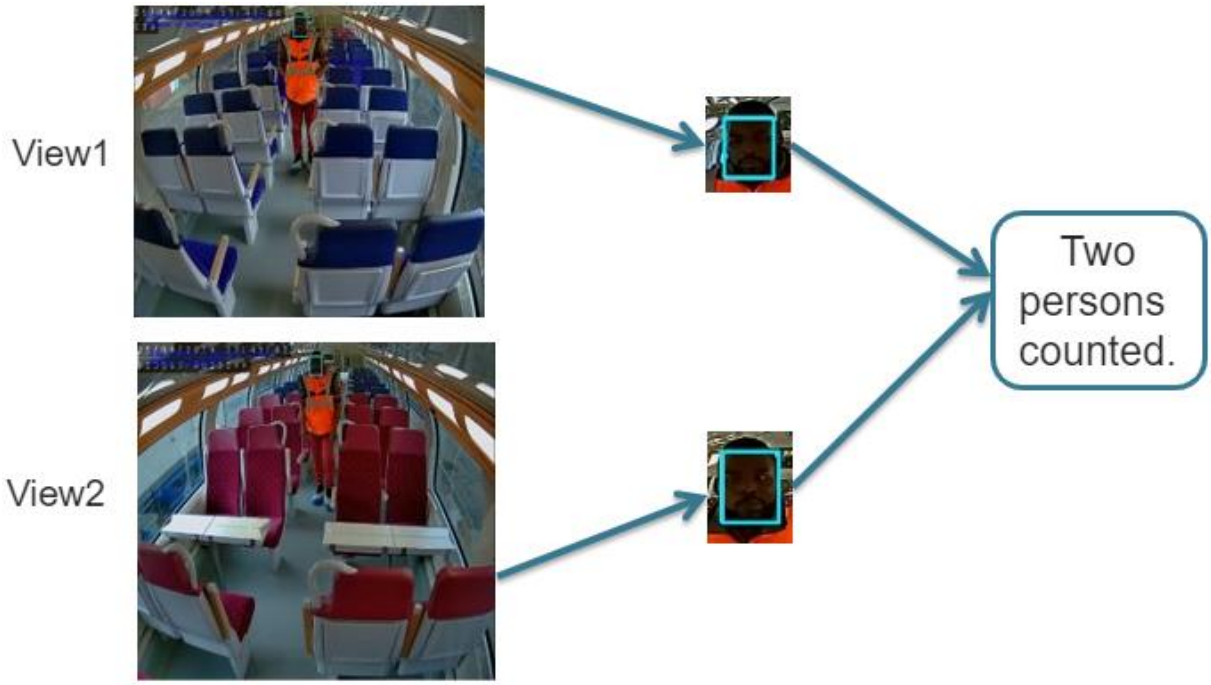

Figure 3: counting with face detection.

3 Results

In this section, we present the experimental details and evaluation results on four datasets: Boss, Sequence, Ardoine 1 and Ardoine 2. We evaluate the performance using the Mean square error (MSE) and the Mean absolute Error (MAE).

$$MAE = \frac{1}{N}\sum_{i}^{N}|C_i - Cr_i| \quad , \quad MSE = \frac{1}{N}\sum_{i=0}^{N}(C_i - Cr_i)^2$$

Where N is the number of test images, Ci and Cri are the estimated and the real count, respectively.

Sequence dataset:
As part of maintenance test of surveillance camera in trainset, a sequence of database has been created. It has 14 videos of one minute with an average of two persons by video.

Ardoine dataset 1:
Contains images recorded from the RER Z2N trainsets during the test of our solution on "les Ardoines" maintenance center near of Paris. It has 50 couple of images with dimensions 704x576 , captured from the two cameras of different trainsets with an average of 4 persons by image.

Ardoine dataset 2 :
Includes 48 recorded videos full HD of dimensions 1920×1080, taken from an IP camera while the same test in "les Ardoines" maintenance center.

	Boss		Sequence		Ardoine 1		Ardoine 2	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MTCNN	2.54	3.36	1.23	2.22	3.12	4.46	1.16	2.13
Haar Cascade	2.76	3.43	1.66	2.87	2.04	6.01	0.78	1.13
Tiny FaceResNet101	**0.16**	**0.28**	**0.12**	**0.14**	**0.55**	**0.68**	**0.23**	**0.23**
MaskRCNN	1.30	2.98	1.32	1.76	1.08	2.05	0.84	1.10
MaskRCNN+PersonReid	0.83	1.75	0.14	0.35	0.98	1.39	0.64	0.92
MakRCNN+ALignedReid	**0.26**	**0.31**	**0.17**	**0.12**	**0.51**	**0.62**	**0.28**	**0.38**

Table 1: Estimation MSE and MAE error on the four datasets.

The results indicate that Tiny-FaceResNet101 and MaskRCNN+AlignedReid have the best performances in terms of MSE and MAE.

As expected, integrating AlignedReid was benefic for the four dataset since it deals with both local and global features, unlike PersonReid.

We have better results using re-identification in counting than only average number of persons segmented in the two frames because of redundancy elimination.

The MAE and MSE have remarkably been reduced in Ardoine 2 because of high resolution of recorded videos.

The performance has been degraded while testing on Ardoine1/2 datasets due to several factors:

- Low quality of images: as shown in figure 4, persons do not appear clearly in the image, so they are not segmented, and hence the face is not detected. Figure 5 provide a comparison between quality of Z2N camera and an IP camera.

Figure 4: impact of low quality of images with Z2N camera.

Figure 5: An example of images with camera IP on the left vs Z2N camera on the right.

- Scale variation: faces and persons so far from camera are not detected and segmented, contrary to those that were close to camera, see figure 6.

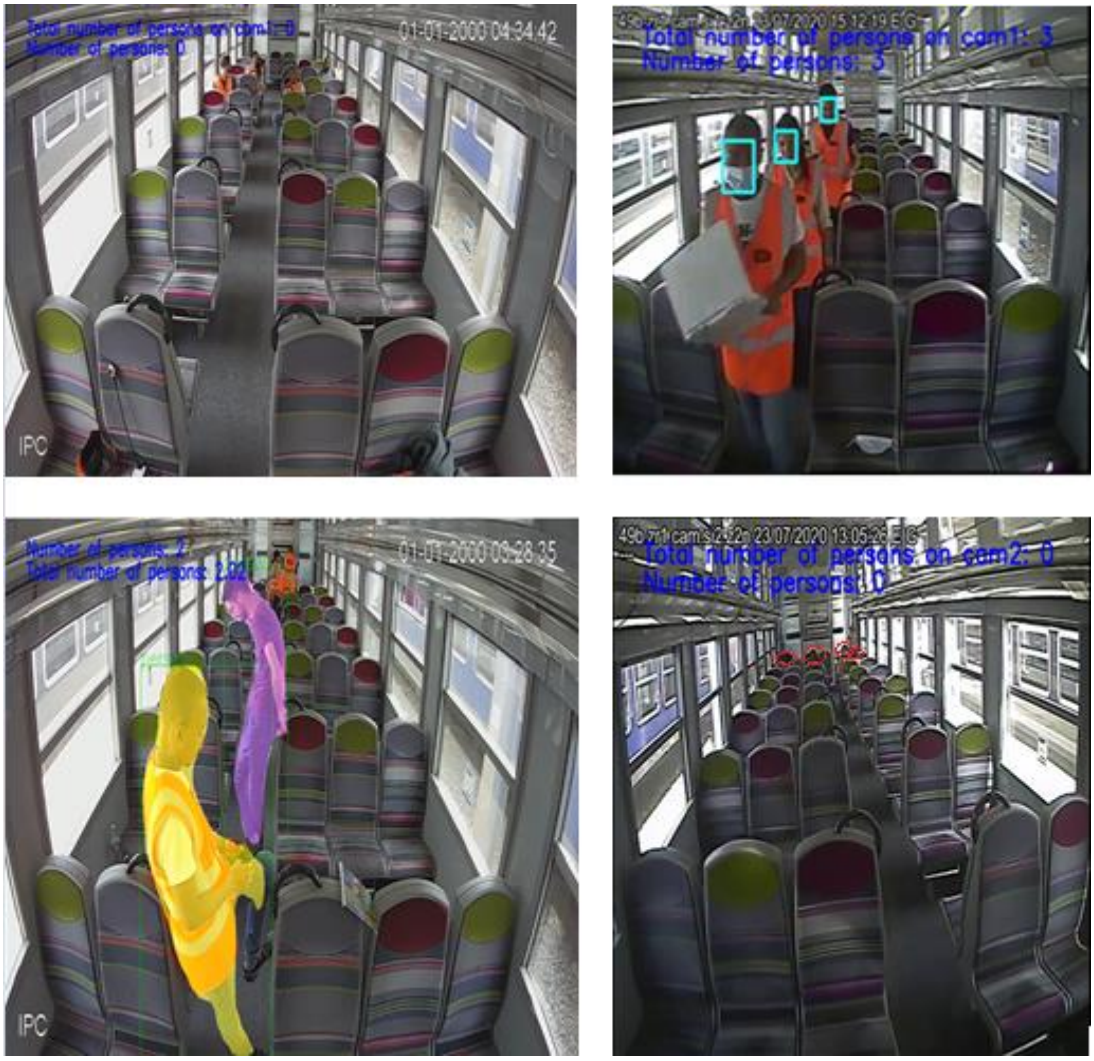

Figure 6: impact of scale variation on counting.

- Position of faces: with Tiny-FaceRes101, in some cases faces that are not totally in front of the camera are not counted. See figure 7.

Figure 7: impact of face position on counting.

4 Conclusions and Contributions

In this paper, we propose two approaches for passenger counting in Z2N trainset which are: counting with MaskRCNN+AlignedReid and Tiny FaceRestNet101. Experiments on multiple datasets have shown that our solution is more suitable for low density counting using high resolution cameras. In order to improve the performance of counting, future work should focus on designing new algorithms for crowd density estimation, and using preprocessing techniques to improve images quality

References

[1] [1] Application of Image Recognition Software at the Platform-Train Interface (S206) RSSB.

[2] [2] Loy et al. (2013) Loy, C.C., Chen, K., Gong, S., Xiang, T., 2013. Crowd counting and profiling: Methodology and evaluation.

[3] [3] J. Liu, C. Gao, D. Meng. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation.

[4] [4] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. University of Washington.

[5] [5] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. Mask R-CNN. Facebook AI Research.

[6] [6] Xuan Zhang1, Hao Luo1, Xing Fan1, Weilai Xiang1, Yixiao Sun1, Qiqi Xiao1, Wei Jiang2, Chi Zhang1, Jian Sun1.AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. Institute of Cyber-Systems and Control, Zhejiang University.

[7] [7] Shuangjie Xu, Yu Cheng, Kang Gu. Jointly Attentive Spatial-Temporal Pooling Networks for Video-based Person Re-Identification.
[8] [8] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks.
[9] [9] R. Lienhart, J. Maydt. An Extended Set of Haar-like Features for Rapid Object Detection.
[10] [10] Peiyun Hu, Deva Ramanan. Finding Tiny Faces. Robotics Institute Carnegie Mellon University.
[11] [11] http://velastin.dynu.com/videodatasets/BOSSdata/