Human-Computer Interaction & Emerging Technologies, Vol. 195, 2025, 44–53
https://doi.org/10.54941/ahfe1006222

Real-Time Object Recognition With Neural Networks in Public Transport – Determining the Utilization of Vehicles Using Existing Camera Systems

Waldemar Titov[1], Julian Knust[2], and Thomas Schlegel[1]

[1]Institute for Intelligent Interactive Ubiquitous Systems (IIIUS), Furtwangen University, Robert-Gerwig-Platz 1, Furtwangen im Schwarzwald, 78120, Germany
[2]Karlsruhe University of Applied Sciences, Moltkestrasse 30, 76133 Karlsruhe, Germany

ABSTRACT

For security reasons, many local public transport vehicles now have cameras installed in their interiors. At present, these can and may only be used in Germany to investigate criminal offenses. Real-time object recognition offers the possibility of counting passengers with existing cameras without saving images or videos. This is not only important for revenue sharing, but can also provide information about bottlenecks when boarding and alighting, or be used to display empty areas in individual carriages of a train. This paper investigates whether object recognition is suitable for determining the utilization of public transport vehicles using individual images from the interior. Test images from the security cameras of a streetcar were analyzed and evaluated with a self-trained Faster R-CNN model. Accuracies of 70% were achieved in the detection of people and free seats.

Keywords: Passenger counting, Artificial neural networks, Innovative public transport

INTRODUCTION

Passenger counting in local public transport is carried out today for many reasons. One of these reasons is the collection of utilization figures for operational planning. Local transport companies receive data on where passengers board and alight and can draw conclusions from this. On the one hand, services can be better adapted to predictable events. On the other hand, service adjustments in daily operations, such as the provision of more or fewer carriages, are possible. This is the approach taken by the transport stock corporation Nürnberg Germany with its counting systems in the subway network. Counting takes place both above the vehicle doors and at the entrances and exits of the platforms (Institut für Verkehrswesen, 2017). Exact passenger numbers are particularly relevant when it comes to revenue distribution in transport associations based on the actual number of passengers transported. Here it is important to know how many people boarded and alighted at a particular stop. The data should be highly accurate and reliable. More and more public transport vehicles are already equipped with security cameras. At the 2016 Conference of Transport Ministers, the ministers recommended equipping public transport vehicles with cameras across the board.

COUNTING PASSENGER

Today, automatic passenger counting is carried out using sensors installed in the entrance areas of the vehicles, among other things. Information on what percentage of the fleet should be equipped with counting devices ranges from 10% (Boyle, 2008; Girshick, 2015) to 100% of the fleet (Chu, 2010). Equipment of less than 100% results in constraints in vehicle deployment planning. It must be ensured that the vehicles with sensors are deployed evenly throughout the network in order to obtain meaningful results. The larger the network, the more difficult it is to meet this requirement. Equipping vehicles with sensors is expensive. The approach presented in this paper uses existing security cameras in the vehicles and uses object recognition to detect people and objects. Sensors above the doors can detect how many people are in the vehicle, but this data can only provide limited information about how the passengers are distributed. This information can be used to inform people at the next stop where there are still free spaces on the next train. Deutsche Bahn presented a project of this kind in 2017. Displays in the door area show the utilization of the train divided into areas. The data comes from sensors mounted above the doors.

There are numerous approaches to using videos for counting passengers. Cheng et al. (2014) used feature-based tracking and trajectory clustering to count passengers boarding and alighting. Escolano et al. (2016) achieved the same with the help of optical flow. Optical flow was also used by Taniguchi et al. (2016) for counting people in crowded environments. Liciotti et al. (2017) present a system with a depth camera (RGB-D). This approach was also followed by Lumentut et al. (2015) and Del Pizzo et al. (2016). Perng et al. (2016) used background subtraction for the recognition of persons. Jaijing et al. (2009) present a method that uses single images. People are counted at a threshold where images are taken vertically from above and the software recognizes the direction of movement based on the position of the face. Existing security cameras in a vehicle are used by Potter et al. (2011), but here too video recordings are used and not individual images.

Liu et al. (2017) used a convolutional neural network for the recognition and spatial-temporal context for tracking passengers' heads in a passenger counting system based on video images. They achieve an accuracy of up to 90%. Similar results were presented by Tome' et al. (2016) presented something similar. They also used a CNN to recognize people and concluded that their method is suitable for the real-time recognition of pedestrians on current hardware.

ARTIFICIAL NEURAL NETWORKS

In terms of structure, artificial neural networks (ANN) are modeled on the human brain. Nodes of the network are also referred to as neurons and store information. ANNs are able to learn from examples and retrieve and apply what they have learned in other situations. ANNs consist of several layers. The neurons can be connected to all neurons of the two neighboring layers (interlayer) and also to each other (intralayer) (Yegnanarayana, 2009). The latter is no longer found in today's leading models. While the initial motivation was to replicate the human brain, in the context of machine learning this term is now understood to mean the abstraction of information processing (Schalkoff, 1997).

Convolutional neural networks (ConvNet or CNN) are specially designed for processing image data. In contrast to conventional ANNs, they can have a three-dimensional structure. The layers do not comprise the entire image, but sub-areas, which in turn can have layers for sub-areas. An image is expected as input; the output can, for example, contain estimates of the predefined class to which the main object of the image belongs (Szegedy et al., 2015; Del Pizzo et al., 2016).

Object recognition can be divided into several areas: Image Classification describes the classification of a photo according to the dominant object in it. The next step, Object Localization, builds on the previous step and makes a prediction about the image area in which the recognized object occurs. Object Recognition classifies and localizes objects contained in the image (Andreopoulos et al., 2013; Chen et al., 2017; Lin et al.,2014).

Although research into object recognition with computers dates back to the 1960s (Andreopoulos and Tsotsos, 2013), the topic has only gained in importance in recent years (Chen et al., 2017), as powerful processors have become available. Deep Convolutional Neural Networks (DCNN) have contributed significantly to this. In 2012, AlexNet (also known as SuperVision, Krizhevsky et al., 2017) won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a competition of algorithms for object recognition, by a wide margin over the other participants. The data set consisted of 100,000 test images with 1000 object categories. In total 8.3 percentage points (Russakovsky et al., 2015) undercut the previous year's result.

In the following year, Girshick et al. (2015) presented the Region-based ConvNet (R-CNN). It generates up to 2000 bounding boxes by grouping similar regions and extracts features using AlexNet. Only the bounding boxes that contain an object are transferred.

In 2013, OverFeat won the localization competition. It is a further development of AlexNet and improves the size accuracy of the bounding boxes (Yegnanarayana, 2009). While R-CNN generates more accurate bounding boxes, the method used by OverFeat is 2 times faster. With Fast R-CNN, the speed differences became smaller (Girshick, 2015). The further development Faster R-CNN uses the results of CNN more efficiently and can thus achieve a further speed advantage (Ren et al., 2015).

EVALUATION

Image material is required to test whether video recordings from the security cameras are suitable for passenger counting. For data protection reasons, no existing recordings could be used. Therefore, these were carried out with voluntary participants in a streetcar at the Karlsruhe public transport depot. They provided a streetcar of the NET 2012 type. It has 11 surveillance cameras. Seven of them point to the interior, 4 each to a door area. The resolution is 704×576 pixels. The recordings are stored on a removable hard disk in a proprietary format. They can be viewed and exported using a program provided by the manufacturer. The 11 cameras for the test recorded a total of 16.5 hours of video material. In addition to the security cameras, two video cameras were installed in the area of the bellows in the middle section for filming.

Figure 1 shows the distribution of the cameras in the track. The triangles indicate the direction in which the cameras are pointing, but do not symbolize the true, significantly greater, viewing distances for reasons of presentation. The cameras only insufficiently cover the area in the two folding bellows. Seats located directly under the camera cannot be seen. Due to the relatively low mounting height in the window area, they can easily be obscured by standing people. As there is no camera on the other side of the vehicle, it is not possible to compare the images. 68 individual images from cameras 3 to 5 and 9 to 11 were used for the training. The evaluation was carried out with 15 images. Overall, 45 people took part in the test. 32 people were given a playing card. This allowed individual groups to be asked to perform actions such as getting on or off the train. In this way, various scenarios could be simulated in the course of the recordings. Towards the end of the images, the test subjects were asked to take off their jackets and hats to prevent the system from only being trained on people in winter clothing.

The Faster R-CNN model was used for the test, as it has the highest accuracy of the available models. For all evaluations, it is possible to specify how many results (in these cases bounding boxes) and the probability from which they should be output. The latter is a measure in percent of how confident the algorithm was in its prediction. The evaluation was carried out with the specification of a maximum of 25 bounding boxes (recognized objects) and a probability of at least 80%.

The Faster R-CNN model (exact name: faster rcnn inception resnet v2 atrous coco) was pre-trained with the COCO dataset (Lin et al., 2014). The dataset contains 200,000 photos with 80 marked objects. The download consists of finished files that can already be used for object recognition. The training can be continued with your own images. Compared to starting from scratch, this means a considerable time advantage.

The pre-trained model was trained with 68 hand-marked video images from cameras 2 to 5 and 9 to 11. In addition, 3 images with free seats from cameras 6 to 8 were included, which only contained markings for free seats. A selection of photos of baby carriages, backpacks and bicycles were also included in the training set. Objects were marked in 48 photos for this part. The training was carried out on a machine with an NVIDIA Tesla P100 graphics card, which significantly accelerates the training and evaluation compared to a CPU The model was trained with five classes: People, seats, baby carriages, bicycles and rucksacks. The status was saved every 1000 steps. After 30,000 steps, no further improvement could be detected and the training was aborted. Screenshots with bounding boxes were taken and evaluated after 0, 10,000 and 30,000 steps.

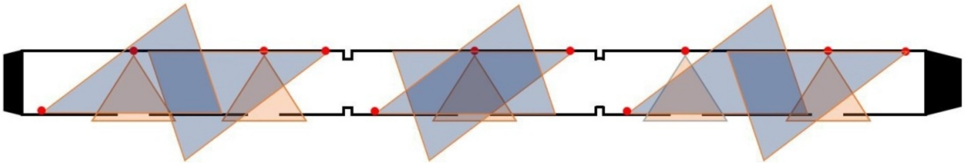

Figure 1: Distribution of cameras in the NET 2012 streetcar.

RESULTS

The following images show a selection of screenshots from the camera images of the track with bounding boxes. The color of the rectangles stands for the class to which the algorithm has assigned the detected object. This is shown on the left above the box, directly behind. It is the probability in percent with which the object corresponds to the class.

The pre-trained construction model is shown on the left-hand side in each case. In the middle are the results of the model with additional training images of cameras 2 to 5 and 9 to 11, as well as images of other objects after 10,000 steps. The right-hand side shows the screenshots after 20,000 further steps. In total, the training was carried out over 30,000 steps.

After 10,000 training steps, more people are recognized and, apart from the folding seats, free seats are recognized with very high accuracy. After a further 30,000 steps, the folding seats are also marked as free seats. However, misclassifications also increase.

The evaluation according to recognized persons in the image is shown for the individual training conditions in the following tables. The order corresponds to Figures 3 to 7. No persons are included in Figure 2. The more objects are included in the image, the worse the results. The base model above all recognizes persons when they are depicted as completely as possible. Faces are not recognized. Free seats cannot be recognized because the model is not trained to do so. There are no incorrect classifications.

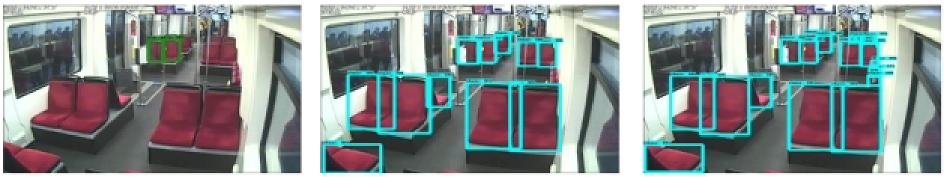

Figure 2: Camera 6 (LTR): basic model, 10,000 training steps, and 30,000 training steps.

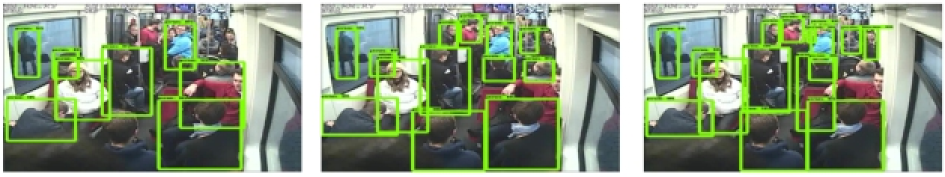

Figure 3: Camera 6 (LTR): basic model, 10,000 training steps, and 30,000 training steps.

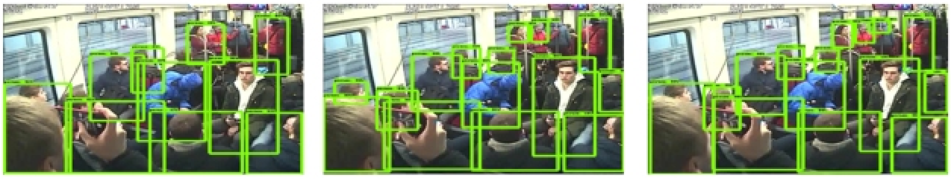

Figure 4: Camera 8 (LTR): basic model, 10,000 training steps, and 30,000 training steps.

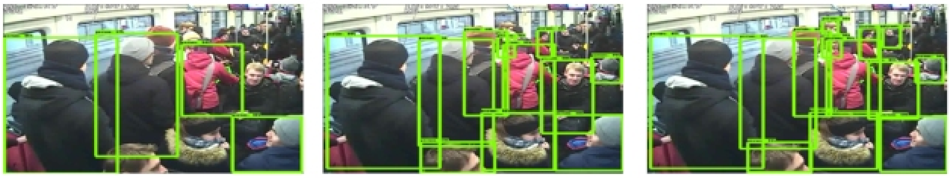

Figure 5: Camera 8 (LTR): basic model, 10,000 training steps, and 30,000 training steps.

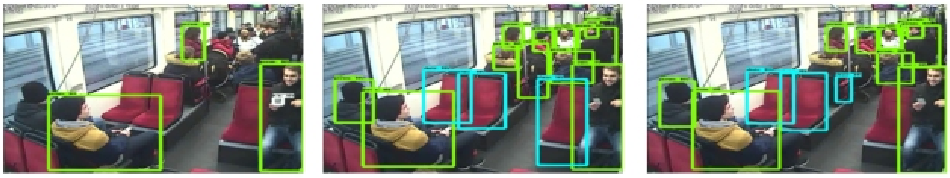

Figure 6: Camera 8 (LTR): basic model, 10,000 training steps, and 30,000 training steps.

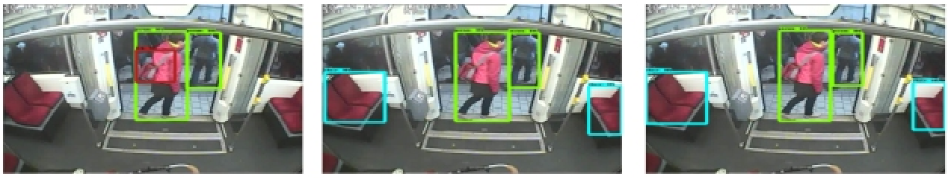

Figure 7: Camera 7 (LTR): basic model, 10,000 training steps, and 30,000 training steps.

After training with marked images from the train, accuracy increases significantly. Now even people whose bodies or faces are partially obscured are recognized. This leads to more people being recognized and a better result. Although backpacks and baby carriages are present in the training set, they are not recognized. A total of 2 people and one rucksack are not correctly classified (incorrectly recognized).

After 30,000 training steps, even more people are marked correctly in individual cases. However, incorrect classifications are now also occurring. Backpacks and baby carriages are still not recognized.

The increase in accuracy in the detection of people and free seats during training can be seen in Figure 8. After the new training with 40,000 steps and 8 images from cameras 6 to 8, people are recognized in a few cases that were not recognized before, but people are also marked twice (for example, the head and also the whole body). This reduces the accuracy.

COUNTING THE DOOR AREAS

Due to the way in which the cameras pointing into the interior are installed, passenger counting is subject to limitations. For this reason, a second test is carried out to investigate how accurately it is possible to detect people in the entrance area. All individual images from a 5-minute video from camera 7 are analyzed using the model with 30,000 training steps and manually checked for unrecognized persons. The result is significantly better with 99.8% of people recognized. Every person was recognized, but in some cases not in every consecutive frame. However, a solution for tracking would still have to be found for productive use. Unlike recordings from the interior, videos are evaluated here in order to capture all passengers. Tracking adds another source of error, which means that the accuracy is likely to deteriorate. The computing power required is very high, as 25 images per second are evaluated. As Figure 9 shows, the baby carriage is also recorded.

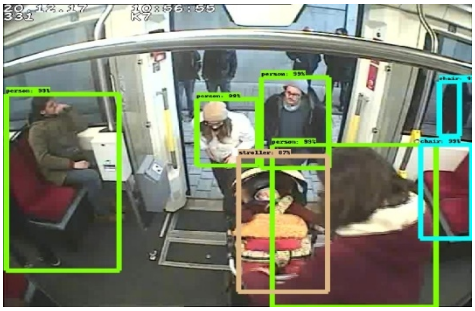

Figure 8: Passenger counting in door area.

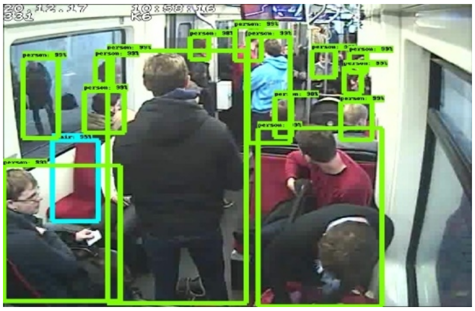

Figure 9: Detection of persons outside tram.

CONCLUSION

This study investigated the extent to which neural networks are suitable for determining the utilization of local public transport vehicles. For this purpose, test recordings were made with the test recordings were made with the safety cameras of a streetcar, a Faster R-CNN model was trained with some of these recordings and the results were evaluated. Improvements were identified during the training process. The accuracy increases significantly with additional training if images from existing recordings are used. Accuracies of around 70% can be achieved with individual images from the cameras of the tested streetcar.

The system presented is not suitable for counting passengers based on which revenue is to be allocated. The results are not accurate enough for this. However, it is possible to record occupancy rates and free seats and display these at the next stop. If videos in the door areas are evaluated, the recognition accuracy increases to 99.8%. Together with a powerful tracking algorithm that has yet to be developed, this would meet the high requirements of an automatic passenger counting system. Another advantage of using this system is the ability to determine the direction in which a passenger is traveling on the train. This means that even with these cameras alone, statements that are more precise can be made about the utilization of the vehicle.

The images were analyzed on a powerful graphics card, which would not be economical to use in a streetcar for this purpose. Much longer computing times would have to be expected there. Depending on the distance between stops, the results may not be available quickly enough. However, this problem can be solved by using faster models with the disadvantage of lower accuracy. This would have to be weighed up on a case-by-case basis. When counting passengers in the door area, videos must be evaluated. In this test, 25 images per second were used. This increases the computing power considerably. One approach would be to reduce the number of frames per second. It would have to be investigated whether this has an effect on the accuracy of detection and tracking.

In conclusion, it should be noted that neural networks are generally able to determine the utilization of vehicles in public transport with individual images from existing cameras. However, the achievable accuracy depends on their mounting position.

ACKNOWLEDGMENT

The authors wish to thank the Karlsruhe public transport company (KVV) for providing a streetcar for the test recordings. Parts of this work is connected to the project regioKArgoTramTrain (Funding ID: RegioInn-24577775), which is funded by the European Regional Development Fund (ERDF) in Baden-Württemberg. We also wish to thank the participants of the user study for their time and valuable input.

REFERENCES

Andreopoulos, A., & Tsotsos, J. K. (2013). 50 years of object recognition: Directions forward. Computer vision and image understanding, 117(8), 827–891.

Boyle, D. K. (2008). Passenger counting systems (No. 77). Transportation Research Board.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834–848.

Cheng, G., Huang, Y., Mirzaei, A., Buckles, B. P., & Yang, H. (2014, June). Video-based automatic transit vehicle ingress/egress counting using trajectory clustering. In 2014 IEEE Intelligent Vehicles Symposium Proceedings (pp. 827–832). IEEE.

Chu, X. (2010). A guidebook for using automatic passenger counter data for national transit database (NTD) reporting (No. NCTR778-03, FDOT BDK85 977-04). National Center for Transit Research (US).

De Potter, P., Kypraios, I., Verstockt, S., Poppe, C., & Van de Walle, R. (2011, September). Automatic available seat counting in public rail transport using wavelets. In Proceedings ELMAR-2011 (pp. 79–83). IEEE.

Del Pizzo, L., Foggia, P., Greco, A., Percannella, G., & Vento, M. (2016). Counting people by RGB or depth overhead cameras. Pattern Recognition Letters, 81, 41–50.

Escolano, C. O., Billones, R. K. C., Sybingco, E., Fillone, A. D., & Dadios, E. P. (2016, November). Passenger demand forecast using optical flow passenger counting system for bus dispatch scheduling. In 2016 IEEE Region 10 Conference (TENCON) (pp. 1875–1878). IEEE.

Girshick, R. (2015). Fast r-cnn in proceedings of the ieee international conference on computer vision (pp. 1440–1448). Piscataway, NJ: IEEE. [Google Scholar], 2.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580–587).

Institut für Verkehrswesen. (2017). Nahverkehrs-Tage 2017: Digital und Disruptiv-Neue Daten und Methoden für einen kundengerechten ÖPNV. Kassel University Press GmbH.

Jaijing, K., Kaewtrakulpong, P., & Siddhichai, S. (2009, May). Object detection and modeling algorithm for automatic visual people counting system. In 2009 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (Vol. 2, pp. 1062–1065). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.

Liciotti, D., Cenci, A., Frontoni, E., Mancini, A., & Zingaretti, P. (2017). An intelligent RGB-D video system for bus passenger counting. In Intelligent Autonomous Systems 14: Proceedings of the 14th International Conference IAS-14 14 (pp. 473–484). Springer International Publishing.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D.,... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13 (pp.740–755). Springer International Publishing.

Liu, G., Yin, Z., Jia, Y., & Xie, Y. (2017). Passenger flow estimation based on convolutional neural network in public transportation system. Knowledge-Based Systems, 123, 102–115.

Lumentut, J. S., Gunawan, F. E., Atmadja, W., & Abbas, B. S. (2015, March). A system for real-time passenger monitoring system for bus rapid transit system. In Asian Conference on Intelligent Information and Database Systems (pp. 398–407). Cham: Springer International Publishing.

Perng, J. W., Wang, T. Y., Hsu, Y. W., & Wu, B. F. (2016, July). The design and implementation of a vision-based people counting system in buses. In 2016 International conference on system science and engineering (ICSSE) (pp. 1–3). IEEE.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S.,... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115, 211–252.

Schalkoff, R. J. (1997). Artificial neural networks. McGraw-Hill Higher Education.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D.,... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1–9).

Taniguchi, Y., Mizushima, M., Hasegawa, G., Nakano, H., & Matsuoka, M. (2016). Counting pedestrians passing through a line in crowded scenes by extracting optical flows. International Information Institute (Tokyo). Information, 19(1), 303.

Tomè, D., Monti, F., Baroffio, L., Bondi, L., Tagliasacchi, M., & Tubaro, S. (2016). Deep convolutional neural networks for pedestrian detection. Signal processing: Image communication, 47, 482–489.

Yegnanarayana, B. (2009). Artificial neural networks. PHI Learning Pvt. Ltd.