A novel smart camera network for real time public transport monitoring and surveillance

Andrea Carboni[1], Giuseppe Riccardo Leone[1], Simone Nardi[2], Alfonso Corrado[2] and Davide Moroni[1]

***Abstract*— In this paper, we present the environment perception layer of the Smart Passenger Center (SPaCe), a novel integrated framework for public transport management. This layer is a pervasive vision architecture for improved safety and security in the context of public transportation. Privacy and technological constraints are still significant limitations for the real-time analysis of video streams from video capture devices installed on public transport vehicles. In fact, in almost all cases, this analysis is carried out offline and lacks any predictive processing, which is now potentially applicable to all transport sectors, thanks to machine learning and artificial vision techniques. The architecture described is designed to combine the output of a set of parallel processing, all running onboard in real-time, thus allowing the separation of the information collected from actual passengers' identities. The analysis highlights aspects that affect travel and travellers safety, such as people's behaviour and the state of maintenance of vehicles.**

I. Introduction

Video analytics is an emerging field of technology that leverages computer vision and machine learning to analyze video feeds and extract useful information automatically. In the context of transit, this technology can be a powerful tool for improving the safety and security of passengers. Indeed, pervasive monitoring of all areas of the vehicle or station is impracticable by human operators. Video analytics can provide a more comprehensive and accurate view of what's happening in the transit system, allowing operators to detect potential safety or security concerns and respond accordingly quickly. For example, unattended bags or suspicious behaviour in a station or on a vehicle can be automatically detected; overcrowding or congestion can represent a safety hazard and can be prevented by monitoring passenger flow. In addition to improving safety and security, video analytics can help transit operators optimize their services and enhance the passenger experience. For example, it is possible to monitor passenger behaviour and detect patterns that can inform the design of new services or improvements to existing services. Overall, video analytics has the potential to revolutionise the way we think about safety and security in public transit. By leveraging the latest technology, transit operators can create safer, more efficient, and more enjoyable passenger experiences. This paper introduces a novel smart camera network that constantly monitors activities in stations, trains, buses and other places of interest. This pervasive vision architecture is the perception layer of the Smart Passenger Center (SPaCe), a fully integrated platform that aims to overcome the complexity of centralized management of public transport infrastructure and vehicles [1]. The SPaCe artificial intelligence engine predicts threats and critical events. It proposes countermeasures by examining the daily flows of people and correlating different data and events, thanks to machine learning and big data analytics [2]. In the next section, some related technologies are referred to. In section III, the design choices, the working pipeline and all the functionalities implemented by the vision system are presented. The experimental setup, the custom model training results and the system's performance are shown in section IV. Section V concludes the paper.

[1]Institute of Information Science and Technologies - National Research Council of Italy, Pisa, Italy
[2]Mermec Engineering srl, Pisa, Italy
Corresponding author: giuseppericcardo.leone@cnr.it

II. Related work

Given the presence of CCTV systems, the application of computer vision to public transit has been investigated for several decades to automatize the analysis of video streams. Early attempts considered problems linked to human presence [3] to assess if a carriage was empty or, if not, counting the number of passengers or empty seats (see e.g. [4] where wavelet transform is used). Evaluation of passenger flows to understand the commuting patterns in multi-modal transport has also received much attention in the framework of smart cities [5]. However, tracking through smart cards (e.g. Oyster card in Transport for London) was soon revealed to provide more accurate and massive data for travel analysis and knowledge discovery. Vision has been therefore applied to address safety and security concerns, such as unattended luggage or suspicious objects [6] and recently for social distancing [7]. Bringing video analytics to extended areas such as stations or to environments made of multiple rooms such as railway and subway cars requires the design of scalable architectures, offering distributed services, e.g. based on edge and fog computing [8], [9]. This brought to the creation of *Smart Camera Networks*, which is a network of sensors where each node has, at the same time, i) vision capabilities for sensing and ii) data processing capabilities for understanding the scene the node perceives. Smart Camera Networks have revolutionized surveillance applications across various domains, such as environmental monitoring, habitat observation, emergency response, disaster recovery, law enforcement, and assisted living. To deploy camera platforms successfully indoors and outdoors, stringent hardware requirements must be met. For outdoor camera networks, power efficiency stands as a crucial prerequisite. Additionally, the effective distribution of available bandwidth among network nodes, especially in dense and large-scale

deployments, is another vital consideration that affects both wired and wireless architectures. Besides efficient hardware utilization, Smart Camera Networks must deliver robust performance while employing energy-efficient strategies to maximize lifespan. In public space surveillance, resource utilization must be optimized to ensure uninterrupted service availability while preserving individual privacy during data collection and handling. A critical requirement is to transmit and store only relevant data, minimizing unnecessary network resource consumption and preventing overwhelming the network interface with excessive and irrelevant information. In this respect, *visual sensor data* demands significantly higher storage and bandwidth compared to *scalar sensor data*. Continuously recording live footage with cameras in outdoor surveillance scenarios proves prohibitively expensive and has very poor scalability features. As a result, many solutions adopt *in node* processing at the sensor nodes, substantially reducing the amount of data that needs to be transmitted and archived [10]. Whether the objective of the surveillance application is event detection or object identification, *in node* data processing aids in preventing false alarms by providing the camera with sufficient information to distinguish genuine events from noise, leaving finer and more accurate processing to other nodes in the network, eventually resorting to *off site* and cloud computing. In the context of ITS, Smart Camera Networks have been applied both to the management of traffic flows and control of railway crossings [11], but also to infrastructure monitoring for safety and security considerations [12]. In these examples, events of interest were mainly linked to vehicles or anomalies not involving images and video streams of persons. For the goal of this paper, instead, analysis of people and related objects in stations and inside carriages is more relevant. A rich corpus of literature is available from video surveillance and Ambient Assisted Living (AAL) communities for these aspects. Significantly, following the recent development in computer vision and the advent of the so-called deep learning, the present work considers state-of-the-art techniques in the field of object detection: among Convolutional Neural Networks (CNNs) with supervised learning models, good accuracy has been reported from R-CNN (Region-Based Convolutional Network) [13], Fast R-CNN (Fast Region-Based Convolutional Network) [14] and Faster R-CNN [15]; regression-based methods like SSD (Single Shot Detector) [16] and YOLO (You Only Look Once) [17] perform better on real-time video processing. Integrating such powerful methods into resource-constrained hardware poses special considerations and ad hoc solutions. Many papers working on this integration concentrate on autonomous navigation [18]. At the same time, applications to large-scale camera networks are still an active research field where there is room, especially for cooperative aspects among the network nodes, e.g. multi-camera tracking, behaviour understanding and re-identification [19].

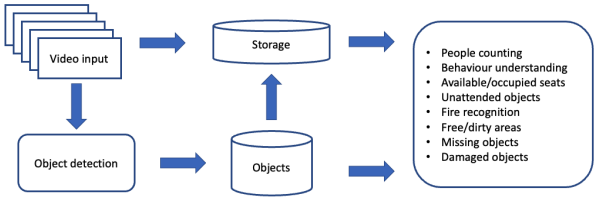


Fig. 1. The object detection task is the core of the processing video pipeline: all the functionalities are built on top of the data structures corresponding to the elements (objects or people) visible in the scene.

III. Materials and Methods

The solution proposed in this paper is based on several computer vision methods orchestrated to provide safety and security services. The main idea is to use an efficient object detector and create customised recognition models trained in public transport scenarios. Therefore the task of object detection is the core of the system; for this research, we chose YOLOv5, which is computationally suitable for real-time also on edge devices like NVIDIA Jetson embedded systems [20]. Figure 1 shows the computer vision processing pipeline: it consists of extracting data from the video stream and constructing a set of data structures corresponding to the elements (objects or people) visible in the scene. The metrics built on top of the object detection task cover two main aspects: the analysis of the passengers and their behaviour and the analysis of the spaces and possibly harmful situations. The functionalities described in the rest of this section are all geared towards passenger safety.

A. Multi camera People Tracking

The system monitors and tracks people's behavior during their travels on public transport. One of the key challenges is tracking individuals across multiple cameras whose field of view is not overlapped, which is referred to as person re-identification task (Re-ID) [19].

To identify tracks belonging to the same person, a clustering algorithm is employed using a multi-modal data fusion system, mainly based on facial and body characteristics (fig. 2). The clustering algorithm utilized in this system is a custom definition of the DenStream [21], which is an extension of the well-known DBSCAN [22] specifically designed to handle streaming data.

Re-identification between different tracks relies on a face feature matching algorithm and is further improved by a body feature matching algorithm when faces are unavailable. Facial biometric features are extracted from cropped images containing the face based on face encoding and recognition frameworks such as Dlib C++ Library [23]. The body's appearance features are associated with visible clothing and accessories, such as the predominant colours of the clothes and the presence of items like trolleys, suitcases, or backpacks. Since encoding faces is not always feasible, the entire person's figure is analyzed, and unique characteristics are extrapolated to construct a robust appearance model of the entire figure. Person feature extractor frameworks are

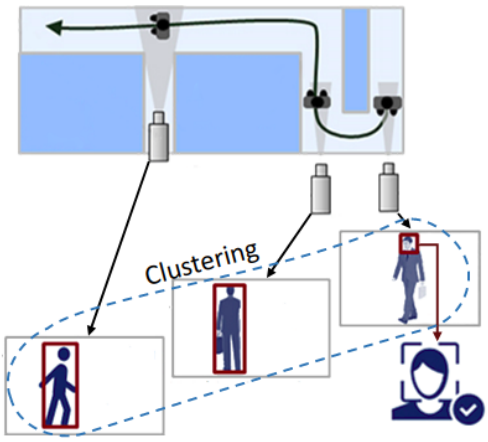


Fig. 2. Multi camera people tracking: for person re-identification, a clustering algorithm is employed using a multi-modal data fusion system, mainly based on facial and body features.

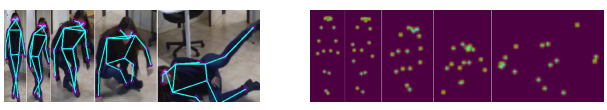
employed in this process, such as FastReID [24] customized with the "Bag of Tricks" method [25], which is a strong baseline in the Re-ID task.

B. Behaviour understanding

Human behaviours are often the cause of many harmful situations. They can be intentional, like a fight or a robbery, or unintentional, like when someone falls on the ground. The human behaviour understanding system is based on a two steps process (fig. 3 and fig. 4): the first one is the human-skeleton extraction which produces a list of coordinate-triplets (x_k, y_k, c_k) (respectively the location and the confidence score of the k-th joint) and the second one is the $3D$ heat-map analysis of the $2D$ skeleton sequences. We represent a $2D$ skeleton as a joint-limb heat-map of size $K \times H \times W$, where K is the number of joints, H and W are the height and width of the frame. The joint-limb heat-map is created by composing K Gaussian maps centred at every joint of the skeleton, where the variance of the k-th Gaussian is related to the c_k confidence score of the relative joint. Finally, a $3D$ heat-map is obtained by stacking all heat-maps along the temporal dimension, which thus has the size of $T \times K \times H \times W$. The resulting $3D$ heat-map is analyzed with a PoseConv3D [26] (a $3D$-CNN based approach) re-trained on a custom data-set of human actions.

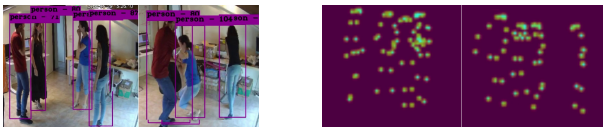
C. People counting

Crowded scenarios could be potentially dangerous for anyone, especially for the elderly and disabled people. It



(a) RGB images and limbs (b) Joints heat-maps

Fig. 3. For every frame of a video, a 2D human skeleton is extracted. Then each heat-map of joints/limbs is stacked along the temporal dimension. Finally, with a 3D-CNN, the 3D heat-map is classified.



(a) RGB images with boxes (b) Global Joints heatmaps

Fig. 4. Although the described process assumes the presence of a single person in each frame, it can easily be extended to the multi-person case, where the k-th Gaussian maps of all people are directly accumulated without enlarging the heatmap.

is essential to know in real-time how many people are in a public vehicle. This information can be used to address the passenger towards the emptier carriages of the train or to advise them to wait for the next bus. The people-counting functionality is generally based on the recognition of the whole person. For indoor and crowded scenarios, like a bus in rush hour, the recognition of the head is a more solid and reliable approach because it is less subject to occlusions (fig. 5-a). This task uses a custom model trained on the CrowdHuman image data-set [27].

D. Seat occupancy

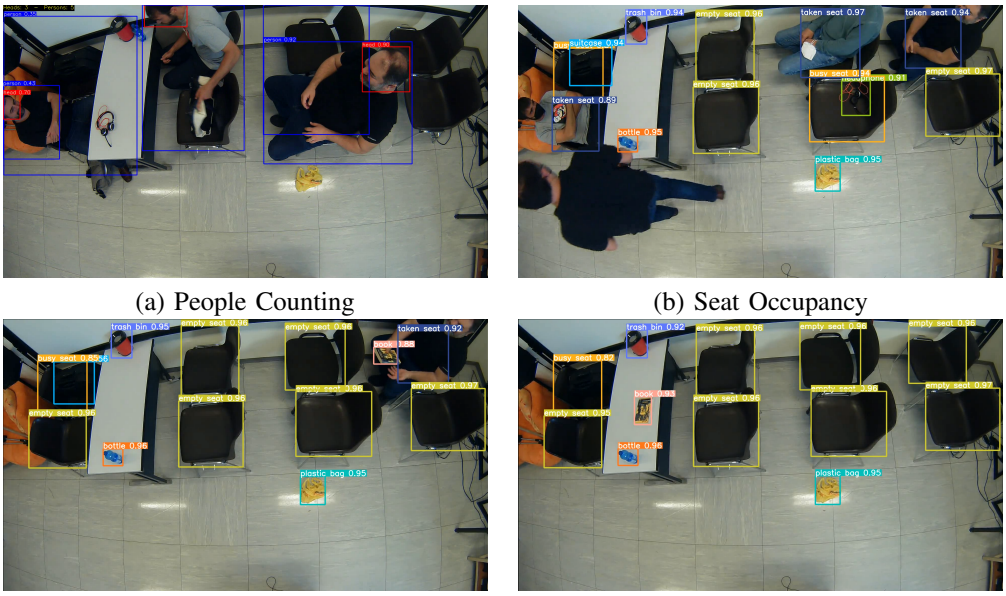
Real-time notification of seat status, especially in long vehicles like trains or metro, can increase passengers' comfort and safety, addressing in advance passengers towards the carriages where seats are available. A combination of two methods is used to determine seat occupancy. The locations of the seats are identified by means of specific regions of interest and compared with the bounding box generated by the people recognition routines, also considering, to improve accuracy, its centre of gravity. In addition, we trained a custom model capable of distinguishing between three different classes: the empty seat, the seat occupied by a person (taken seat) and the seat occupied by an object (busy seat) (fig. 5-b). This third option is very useful for discovering an unattended object.

E. Unattended and lost objects

In the last decades, unattended objects in public spaces are always considered a potentially dangerous situation. In our system, to determine if an object is to be considered unattended, the output of the object detector is monitored. If no person is present near the recognised object or interacts with it within a predefined time slot, it is marked as unattended. This approach is refined using the output of the classifier trained with a custom model that can distinguish between a person, an object or a seat. If the object detector fails to recognise an object left on a seat marked as "busy" by the custom model, it is possible to report it. As an example, in fig. 5-c a suitcase is recognized in the top left seat: after a while (fig. 5-d), the recognition fails, but the seat is still in the busy status, thus indicating that something is on the seat. The unattended object is classified as "lost" if it is on the "white list" of not dangerous items, like books or glasses.

F. Periodic check for trash and damages

Hygiene level, especially after the mandatory pandemic routines, is fundamental to perceiving a safe environment.



(a) People Counting (b) Seat Occupancy

(c) Unattended object (d) Trash and Lost items

Fig. 5. Functionalities based on YOLOv5 object detection: (a) people counting is based on the recognition of the head which is more reliable than person detection in crowded scenarios; (b) for the seat occupancy three status are detected: available (empty seat), taken by a person (taken seat), occupied by an object (busy seat); (c) if no person is near a detected object for a predefined amount of time this is marked as unattended; (d) objects in the "trash list", like empty bottles, plastic bags or papers on the floor, are reported in real-time to the cleaning staff.

Real-time detection of trash and dirty areas can contribute to the efficiency of the cleaning staff and thus to a better travel experience. Objects in the "trash list", like empty bottles, plastic bags or papers on the floor, are reported as soon as they are detected. See (fig. 5-c and fig. 5-d) Another periodic check is done at the end of the journey to find any anomalies, such as damages or missing items. When the vehicle is not moving and there are no passengers in it, the checkpoint image is compared with a reference background image which typically represents the clean vehicle at the start of the journey. Environmental conditions, especially lighting changes, can add noise to the image obtained by subtraction: in these cases, an erosion filter is applied to remove the noise. With this image processing step, it is possible to detect any areas that show significant differences.

G. Fire detection

There is no need to explain why fire is always dangerous, especially in public transport vehicles, where also the use of lighters is forbidden. For this functionality, we used the custom model [28]; the data set has been provided by the repository and contains samples of fire and smoke datasets from Kaggle [29].

IV. Experimental Activities

Although there is an agreement with some local transportation stakeholders, we did not have the chance to test the system on real scenarios like trains or buses, yet. Because of this impossibility at present for the experimental phases, we report in this section some early results collected in a simulated environment. The experimental setup comprises eight seats and a table to match a possible scenario inside a railway carriage. (Fig.6). Some volunteers have lent themselves to simulating movements compatible with those of a traveller, including the presence of bags, backpacks, books, waste, etc., to meet the requirements imposed by the definition of the algorithms. In front of the entrance area is positioned the Facecam camera, which is useful for facial recognition. The two cameras, Frontcam and Roofcam, monitor the seats respectively lengthwise and from above.

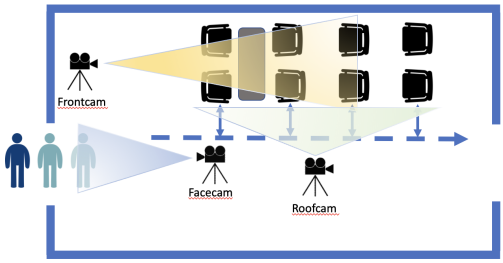


Fig. 6. The experimental setup replicates a possible scenario inside a railway carriage. Three cameras are used: the Facecam camera, which is useful for facial recognition, is positioned in front of the while the other two cameras, Frontcam and Roofcam, monitor the seats respectively lengthwise and from above.

A. YOLOv5 custom model

To reach high accuracy in the object detection task, needed for the functionalities presented in Section III, a custom model has been created using the training feature of YOLOv5 software. Using video frames of the reference scenario (taken from the Roofcam, the camera from above), we selected 13 classes of interest for a total of 916 instances that have been manually annotated. Due to the relatively small number of

Class	Instances	Precision	Recall	mAP50
all	916	0.992	0.739	0.777
backpack	36	0.977	0.944	0.946
book	42	1	0.888	0.941
bottle	62	0.987	0.952	0.957
busy seat	77	0.987	0.985	0.987
empty seat	283	0.996	0.999	0.993
handbag	3	1	0	0.335
headphone	35	0.99	1	0.995
plastic bag	56	0.99	0.982	0.995
smartphone	16	0.968	0.688	0.793
suitcase	61	1	0.925	0.954
taken seat	175	1	0.999	0.985
trash bin	66	0.996	0.985	0.985
umbrella	3	1	0	0.0007

TABLE I

THE CLASSES OF THE CUSTOM MODEL AND THEIR METRICS

instances, no improvement was observed after 382 iterations of the training phase. Figure 7 shows how, after this epoch, the value of precision, recall and mean Average Precision (mAP) are practically constant.

Table I shows the details of the annotated images divided by class. In the first column are the classes' names: ten of these are everyday items present on a train during the travel, i.e. suitcase, bottle or book. Three unique classes, the most numerous, represent the seat occupancy status: they mean to distinguish if there is nothing on a seat (empty), if there is a person (taken) or if there is an object (busy).

B. Performance metrics

it is useful to remember how the parameters in table I are calculated:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where FP, FN are the errors (False Positives and False Negatives) and TP are the good predictions (True Positives). Precision is a measure of "how often the model guesses correctly" while Recall is inversely proportional to "how often the model misses a correct guess". The overall precision of the detection, which is 99.2%, is very satisfying. As a downside, the overall recall, which is influenced by the false negatives, is 73.9%.

For an overall result that takes into account the total number of errors the reference is Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP}{TP + FP + FN} \tag{3}$$

The third part of the equation 3 is obtained because TN, in the case of object detection, is not taken into account. We can interpret it as all correctly undetected objects (i.e. background). The confusion matrix in fig. 8 shows that for almost all the valid classes accuracy is over 90% except for "smartphone" which is 75%. We do not consider the classes "umbrella" and "handbag" because of their under-representation. The accuracy of the object detector is directly connected with some routines of the system and it is a starting point for other functionalities. In particular:

- seat occupancy routine is connected with the accuracy values of the classes "taken seat", "empty seat", and "busy seat" and therefore is an outstanding 99%
- unattended and lost objects recognition rely on the "busy seat" detection which is an indicator that something is present on that seat. This accuracy is 97%
- trash detection routine accuracy depends on the trash object list... in this first model only the classes "bottle" (97%) and "plastic bag" (98%) can be considered as trash, but the list should be increase in future refinements

The other routines, which do not rely on the custom model, show good accuracy values in this preliminary test scenario:

- people counting routine, based on head recognition, has an error rate of 4.26% in environments with few people and rises to 11.91% in crowded environments
- the baseline used in the multi-camera people tracking routine can reach 94.5% rank1 accuracy and 85.9% mAP on Market1501 [30]

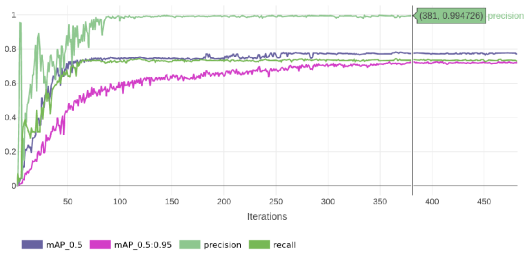


Fig. 7. Metrics of the YOLOv5 custom model.

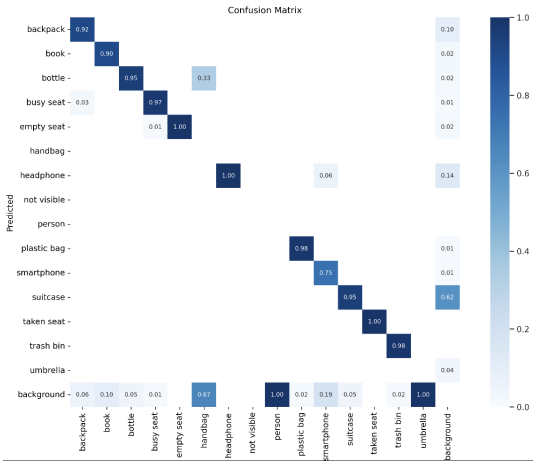


Fig. 8. Confusion matrix of the YOLOv5 custom model.

- human behaviour understanding routines, based on skeleton extraction and $3D$-CNN, have an error rate of 8.30% in environments with few people and rise to 11.7% in crowded environments
- for fire recognition, due to the intrinsic danger of generating fire and smoke in indoor environments, only resources available online have been used. On sample video of indoor scene the measured error rate is 8.79%

V. Conclusions and future work

In this paper, we presented the design, development and prototype testing of a pervasive vision architecture for improved safety and security in the context of public transport. The use of a custom model has proved crucial to reach high accuracy in the object detection task needed for the functionalities presented. The novelty of this model is the ability to determine if a seat is occupied by a person or an object. This information is also used to double-check for unattended or lost items at the end of the journey, improving the accuracy of these detections.

Despite the impossibility of using real scenarios at present for the experimental phases, the early results of the computer vision routines show an acceptable error rate and good robustness that makes them ideal candidates for more extensive experimentation. More testing of the multi-camera configuration is needed for a thorough monitoring of carriages. In the close future, in order to make the custom model available for the scientific community, we plan to extend the data-sets increasing the number of items and the variability of the public transport scenarios: beyond trains and buses also the indoors of stations, waiting rooms and even outdoor waiting areas could be possible places where the presented system could hopefully improve passengers' safety and comfort.

References

[1] A. Corrado, S. Barba, I. Carozzo, and S. Nardi, "Smart passenger center: Real-time optimization of urban public transport," *The Int. FLAIRS Conf. Proc.*, vol. 36, no. 1, May 2023. [Online]. Available: https://journals.flvc.org/FLAIRS/article/view/133300

[2] G. R. Leone, A. Carboni, S. Nardi, and D. Moroni, "Toward pervasive computer vision for intelligent transport system," in *2022 IEEE Int. Conf. on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2022, pp. 26–29.

[3] F. Buemi, M. Esposito, F. Flammini, N. Mazzocca, C. Pragliola, and M. Spirito, "Empty vehicle detection with video analytics," in *Image Analysis and Processing–ICIAP 2013: 17th Int. Conf., Naples, Italy, September 9-13, 2013, Proceedings, Part II 17*. Springer, 2013, pp. 731–739.

[4] P. De Potter, I. Kypraios, S. Verstockt, C. Poppe, and R. Van de Walle, "Automatic passengers counting in public rail transport using wavelets," *automatika*, vol. 53, no. 4, pp. 321–334, 2012.

[5] A. Rocha Neto, T. P. Silva, T. Batista, F. C. Delicato, P. F. Pires, and F. Lopes, "Leveraging edge intelligence for video analytics in smart city applications," *Information*, vol. 12, no. 1, p. 14, 2020.

[6] Hanavi and F. Hidayat, "Intelligent video analytic for suspicious object detection : A systematic review," in *2020 International Conference on ICT for Smart Society (ICISS)*, 2020, pp. 1–8.

[7] K. Gautam *et al.*, "Video analytics based intelligent transport system for passenger flow forecast and social distancing indication," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 7, pp. 2709–2721, 2021.

[8] G. R. Leone, D. Moroni, G. Pieri, M. Petracca, O. Salvetti, A. Azzarà, and F. Marino, "An intelligent cooperative visual sensor network for urban mobility," *Sensors*, vol. 17, no. 11, p. 2588, 2017.

[9] D. Moroni, G. Pieri, G. R. Leone, and M. Tampucci, "Smart cities monitoring through wireless smart cameras," in *Proc. of the 2nd Int. Conf. on Applications of Intelligent Systems*, 2019, pp. 1–6.

[10] M. Magrini, D. Moroni, G. Pieri, O. Salvetti, A. Perallos, U. Hernandez-Jayo, E. Onieva, and I. García-Zuazola, "Smart cameras for its in urban environment," *Intelligent Transport Systems: Technologies and Applications*, pp. 167–188, 2015.

[11] M. Magrini, D. Moroni, G. Palazzese, G. Pieri, O. Salvetti, D. Azzarelli, and A. Spada, "An infrastructure for integrated management of urban railway crossing areas," in *2015 IEEE 18th Int. Conf. on Intelligent Transportation Systems*. IEEE, 2015, pp. 42–47.

[12] G. R. Leone, M. Magrini, D. Moroni, G. Pieri, O. Salvetti, and M. Tampucci, "A smart device for monitoring railway tracks in remote areas," in *2016 Int. Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*. IEEE, 2016, pp. 1–5.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the int. conf. on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 580–587.

[14] R. Girshick, "Fast R-CNN," in *Proc. of the int. conf. on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 1440–1448.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th Europ. Conf., Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the int. onf. on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 779–788.

[18] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[19] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[20] NVidia, "Nvidia jetson embedded systems," Last retrieved on 28.05.2023. [Online]. Available: https://www.nvidia.com/it-it/autonomous-machines/embedded-systems/

[21] F. Cao, M. Estert, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in *Proc. of the 2006 SIAM int. conf. on data mining*. SIAM, 2006, pp. 328–339.

[22] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, 1996, pp. 226–231.

[23] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[24] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for general instance re-identification," *arXiv preprint arXiv:2006.02631*, 2020.

[25] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. of the int. conf. on Computer Vision and Pattern Recognition workshops*. IEEE, 2019.

[26] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. of int. conf. on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[27] Ultralytics, "Yolov5 github repository," Last retrieved on 28.05.2023. [Online]. Available: https://github.com/ultralytics/yolov5

[28] Spacewalk, "Yolov5 fire detector github repository," Last retrieved on 28.05.2023. [Online]. Available: https://github.com/spacewalk01/yolov5-fire-detection

[29] Kaggle, "Kaggle online community platform," Last retrieved on 28.05.2023. [Online]. Available: https://www.kaggle.com

[30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. of the Int. Conf. on Computer Vision*. IEEE, 2015.