Toward Pervasive Computer Vision for Intelligent Transport System

Giuseppe Riccardo Leone*, Andrea Carboni*, Simone Nardi†, and Davide Moroni*
*Institute of Information Science and Technologies
National Research Council of Italy, Pisa, Italy
Email: *name.surname*@isti.cnr.it
†Computational Science Department
Eikontech srl, Pisa, Italy
Email: simone.nardi@eikontech.it

***Abstract*—The Smart Passenger Center (SPaCe) is a fully integrated platform that aims to overcome the complexity of centralized management of public transport infrastructure and vehicles. The SPaCe artificial intelligence engine predicts threats and critical events and proposes countermeasures by examining the daily flows of people and correlating different data and events, thanks to machine learning and big data analytics. All this massive data comes from a pervasive smart camera network that constantly monitors activities in stations, trains, buses and other places of interest. In this work, we present the idea of this computer vision distributed sub-system, the state of the art of the techniques involved and the advanced functionalities that this intelligent surveillance system offers to the upper layers. Everything is developed following the privacy-by-design paradigm; namely, no real image is recorded or transmitted, but all the elaborations take place on the edge nodes of the system.**

***Index Terms*—Intelligent Transport System, pervasive computing, edge computing, privacy-by-design, multiple people tracking**

I. Introduction

With public transport (trains, planes, buses, trams and ships) increasingly accessible to a more significant number of people, public transport operators are facing the dual challenge of meeting the growing demand for ever-increasing passenger flows by ensuring at the same time high levels of quality and security. To meet this challenge, infrastructures and vehicles must be safe and protected from different points of view. Today, areas of interest are fully monitored with closed-loop video systems. However, security operators must monitor and verify the video stream, around the clock, to identify anomalies and manually generate alarms.

Currently available systems, such as the one already present in the public transport vehicles, are used only for security purposes and only by law enforcement agencies. However, the potential of a closed-circuit control system for confined environments, such as a vehicle, is manifold if analyzed through modern automatic detection and classification systems. A modern video surveillance system should be designed to allow transport service operators to prevent, manage and post-analyze unexpected events.

This work is supported by Regione Toscana grant POR-FESR 2014-2020 axe 1 action 1.1.5, CUP: 3647.04032020.157000059 SPaCe

To the author best knowledge, there are two most important examples in this sector: the NAIA system [1], and the MASTRIA system [2]. Both systems are considered experimental solutions not yet tested in the real environment. Those solutions provide suggestions on the state of the environment by examining the flows of people in transit in critical infrastructures. The objectives of such systems are: optimize public transport assets, react to incidents across a city, incorporate all stakeholders in decision-making and execution, deliver smoother journeys and predict and adapt to ridership variations. The Smart Passenger Center (SPaCe) system [3], that we present, allows public transport operators to overcome the complexity of centralized management of infrastructure and vehicles and to meet the safety and security requirements of all interested parties. Thanks to the artificial intelligence engine of the SPaCe integration platform, service managers will be able to improve the efficiency of their operations by drastically reducing the impact of security events. Furthermore, SPaCe enables significant cost savings by providing operators with the ability to focus on critical events rather than micro-management. Moreover, SPaCe enables the transition from the traditional subsystem approach, where a dedicated team is needed to constantly monitor activities, towards a smarter integrated approach that proactively treats an increasing volume of information available in security control rooms. The new platform, which correlates different data and events by crossing them thanks to machine learning and data mining, makes it possible to predict threats with discernment and propose countermeasures. Thanks to SPaCe, it is possible to reduce the costs of personnel and vehicles, optimizing their

Fig. 1. Privacy by design: the video streams are processed on edge and no image is stored permanently in all operations involving sensitive personal data such as seat counts or passenger analytics

presence only in places and moments of actual need. SPaCe allows passengers to reduce delays caused by unforeseen events, allowing a predictive variation of passenger flows. In this work, we focus on the SPaCe computer vision subsystem that constantly monitors activities in stations, trains, buses and other places of interest: it is a pervasive smart camera network whose main task is people counting and passenger's journey tracking; it also supports the service operators with a number of event notifications such as unattended objects or damage detection, smoke and fire alarm and dirt and waste report. In the following sections, we present the state of the art of basic computer vision techniques involved and the advanced functionalities that the system offers to the upper layers.

II. System design and functionalities

The system architecture follows the privacy-by-design paradigm: no real image is recorded or transmitted, but all the elaborations take place on the system's edges. Therefore we focus on the computer vision pipeline used on every smart sensor which is equipped with a stereo camera and a powerful NVIDIA Jetson AGX Xavier [4]; this AI embedded device has the computation and memory capabilities to run all the expected elaborations. As you can see in fig. 2 the first step is to classify and locate all elements in the image with *object detection*, which is useful to locate people and items in the scene. For passengers identification we relay on *face recognition* which is also useful to build the person model used in *people tracking*. For better estimations and measures, we use stereo cameras or depth sensors in combination with standard colour cameras. Let us see a brief review of the fundamental computer vision techniques heavily used in the images processing pipeline.

Object detection: The best results for this task are obtained with convolutional neural networks (CNNs) with supervised learning models. R-CNN (Region-Based Convolutional Network) [5] allows to locate objects by training a model using a small amount of annotated data; SPP-net (Spatial Pyramid Pooling) [6] aims to eliminate the need to provide fixed-size images to the classifier and therefore to avoid cropping or resizing of the images; Fast R-CNN (Fast Region-Based Convolutional Network) [7] guarantees better mean average precision than the previous ones with single-stage learning. Faster R-CNN [8] generates region proposals in a less expensive way than both R-CNN and Fast R-CNN. R-FCN (Region-based Fully Convolutional Network) [9] is completely convolutional, and unlike other RPN-based methods, it replaces FC layers after ROI pooling with simple and low-cost computational operations. SSD (Single Shot Detector) [10] is an object detection method that makes use of a single deep neural network, ensuring a speed suitable for real-time video processing. YOLO (You Only Look Once) [11] as SSD is a regression-based method, which is computationally suitable for real-time (can even reach 155 fps) but behaves worse than SSD on small objects or near objects among them.

Face recognition: The main technique to identifying people is based on Facial Recognition (FR) [16]; more specifically

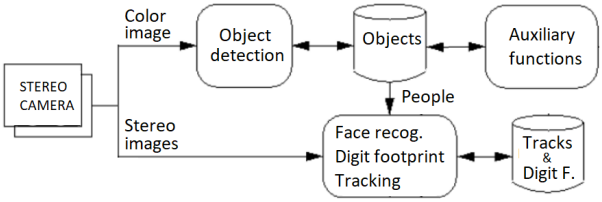

Fig. 2. All the elaborations take place on the edge nodes: after the object detection task the system performs face detection, tracking and counting with people and auxiliary functions with other items; these auxiliary functionalities are related to the detection of dirt and waste, unattended objects, damaged or missing equipment, smoke and fire.

in the SPaCe system we are interested in Face Identification: given an image of a face, a numerical representation is generated, which is used to search within a database if the subject is known. The Deep Learning algorithms used for FR are generally trained through large datasets such as LFW [17] and VGGFace2 [18]. There are many problems with the FR: quantization errors [19], unbalancing of the classes in the dataset [20], [21] and, above all, different resolutions [22]. The latter is because the datasets used in the training phase are composed of high resolution images while, in surveillance systems, you can have images with very low resolutions (up to 8x8 pixels). To face this problem we find the SuperResolution techniques [23], [24], projections in a common space [25] and Cross-Resolution training [22].

People tracking: Single-camera people tracking means assigning an identification number (ID) to all the people present in a given frame and recognizing the same subjects in the following frames by carrying forward the assigned IDs. If someone is no longer present, the ID is inserted in the list of *disappeared*; every time a new person appears, a new identification is assigned only after checking if the person is not in the list of *missing*. The problem just described goes under the name of person re-identification, and many approaches have been proposed in the literature [26]. Deep Learning has made it possible to obtain great performances also in tracking [27], [28]. A reference site that shows the results of the best algorithms is [29]: the winner of the CVPR19 competition proposed in [30] is still the top algorithm today. It uses a detection based on Faster R-CNN and a similarity estimation based on features of size 128 produced by the patches of the converted image in the HSV color space. In the second place, there is [31] whose code in open-source format is freely usable and modifiable.

Relying on the previous computer vision techniques, the system implements both people-based and auxiliary functionalities. The first set involves people counting and multi-camera passenger tracking, while the auxiliary one refers to the detection of "dirt and waste", "unattended objects", "damage or missing equipment" and "smoke and fire". Recent related works for these tasks will be presented in the following paragraphs.

People counting: A first implementation of this functionality is obtained by counting the instances of people found

by an object detector. This approach works well for small to medium-sized, uncrowded spaces where people are distinct. In the case of large and crowded places, on the other hand, techniques based on the estimation of the density of [32], [33] instances, achieves higher performance than the approach described above. Recently, methods based on self-supervised and unsupervised strategies have been proposed in the literature. In particular, in [34] the authors used a self-supervised approach exploiting the *learning-to-rank* paradigm, while in [35] a technique based on the *winner-takes -all* applied to the representations produced by a DCNN giving life to an unsupervised framework. Differently, the authors in [36] have proposed a self-supervised approach based on the idea that the count can be decomposed along with the spatial components of the image by acting independently on their partitions. In [37] an adversarial learning approach is used to reduce the blurring effects typical of the estimation methods using density maps, again using a loss that guarantees spatial consistency of the count. Recently, some papers such as [38]–[40] have successfully attempted to combine regression approaches with detection approaches in an attempt to make the most of the advantages of both methodologies.

Garbage, damages, unattended and missing objects: An efficient and precise object detector is the basis for the implementation of all these functionalities in conjunction with other space-time information (fig. 3). As far as the identification of dirt or abandoned objects is concerned, these checks will be made only in precise moments or at the end of the trip and in the absence of people; in this context, the correct classification is already the desired result. Damage can also be more easily identified when the vehicle is empty using a reference image. Implementation of these functions while passengers are on board will be studied.

Smoke and fire detection: A comprehensive survey of traditional computer vision methods, including some considerations on detection using multi-view and multimodal systems, is presented in the book [12]. Methods based on deep learning have proven to be good in detecting fires, especially to prevent the limitations of traditional methods due to the high false alarm rate and low precision in distinguishing smoke from the isolated haze. For a recent and complete review, we refer to [13], [14]. In [15] a new deep network with normalization (DNCNN) with 14 layers is proposed to implement automatic feature extraction and classification. The experimental results show that the proposed method achieves false alarm rates

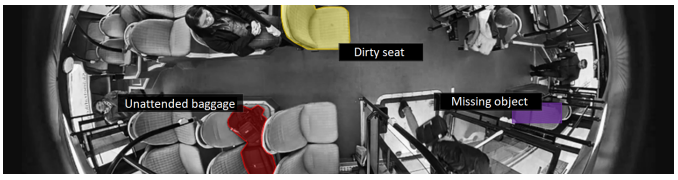

Fig. 3. Examples of auxiliary functionalities: supporting cleaning activities, reporting abandoned objects, identifying damage and vandalism, detecting fire or smoke.

Fig. 4. Digital footprint: It is a mathematical model associated with each person based on the face, physical characteristics and appearance; passenger recognition is based on this information every time it is seen by a camera.

lower than 0.60% and detection rates higher than 96.37% on a reference dataset not freely available. Some datasets useful for benchmarking smoke detection algorithms are [41] and [42].

Privacy-by-design passengers identification and tracking: The SPaCe system should be able to reconstruct the complete route taken by people during their daily journeys, from the departure to the arrival. Therefore, the system should implement a real-time tracking process of people through the network of cameras mounted on the vehicles and in the parking and transit areas. All this must be achieved in compliance with current legislation on privacy, i.e., no image with recognizable people must be stored on permanent media nor transmitted on the network. Light variations, crowded environments and the consequent partial or total occlusions, different views and angles of the cameras make this task particularly difficult compared to a small indoor space. To compare people in different frames of different sensors, we need an appearance model that allows for effective correlation. With a good facial image, the phenotype characteristics of the passenger such as gender and age can be inferred; the face, however, is not always clearly visible; it is necessary to consider the entire figure by extracting unique characteristics to create a robust appearance model: from the analysis of the silhouette the height and build of the person observed can be estimated, and appearance characteristics associated with clothing or accessories can also be taken into consideration (e.g., colour of the clothes, glasses, trolley, suitcases, backpacks). Based on the face, physical characteristics and appearance, the SPaCe system builds a Digital Footprint (DF) of the observed person, that is a mathematical model associated with that person, which is helpful for its recognition (re-identification) every time it is seen by a camera (fig. 4). The probability of success of this recognition is greatly influenced by the clarity of the image, the shot's perspective, and any occlusions present. In favourable frontal framing situations, the system guarantees a clear identification of the people observed.

An interesting starting point for the construction of a fast and reliable digital footprint is the one based on the *deep features* that we find described in [28]. To calculate these characteristics, a model trained on millions of human images is used; the main limitation of this approach is the case of a very large bounding box that includes too much background or information not relating to the person.

III. Conclusion

In this paper, we introduce the Smart Passenger Center (SPaCe), an integration platform for managing and supporting the operators of public transport systems. We focused on the

Computer Vision Subsystem of SPaCe which leverages edge computing and smart cameras to offer advanced services in a robust, scalable and privacy-preserving way. Still in the design phase, the work in progress will explore the interplay between computer vision and pervasive technologies and will address challenging aspects related to inter-node communication and orchestration. Results will be demonstrated in a realistic experimental setup using an installation on test carriages to cope with and tackle the actual issues of the real world.

REFERENCES

[1] Thalesgroup. Tha NAIA homepage. https://bit.ly/naia_thalesgroup
[2] Alstom. The MASTRIA homepage. https://bit.ly/mastria_alstom
[3] Mermec. The SPaCe project. https://bit.ly/mermec_space
[4] NVIDIA. Jetson Embedded Systems. https://bit.ly/NvidiaXavier
[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", IEEE Conf. on Computer Vision and Pattern Recognition, pp. 580-587, 2014
[6] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 2015
[7] R. Girshick, "Fast R-CNN", 2015 IEEE/CVF Int. Conf. on Computer Vision, pp. 1440-1448, 2015
[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, June 2017
[9] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks", 30th Int. Conf. on Neural Information Processing Systems (NIPS'16), ACM, 2016
[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A.C. Berg, "SSD: Single Shot MultiBox Detector", 14th Europ. Conf. on Computer Vision, Springer LNCS, vol 9905, 2016
[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 779-788, 2016
[12] A. E. Cetin, B. Merci, O. Günay, B.U. Töreyin, and S. Verstockt. Methods and techniques for fire detection: signal, image and video processing perspectives. Academic Press, 2016
[13] A. Gaur, A. Singh, A. Kumar, A. Kumar, and K. Kapoor, "Video Flame and Smoke Based Fire Detection Algorithms: A Literature Review", Fire Technology 56, pp. 1943–1980, 2020
[14] S. Khan, K. Muhammad, T. Hussain, J. Del Ser, F. Cuzzolin, S. Bhattacharyya, et al., "DeepSmoke: Deep Learning Model for Smoke Detection and Segmentation in Outdoor Environments", Expert Systems with Applications, Vol. 182, Springer, Nov 2021
[15] Z. Yin, B. Wan, F. Yuan, X. Xia, and J. Shi, "A Deep Normalization and Convolutional Neural Network for Image Smoke Detection", IEEE Access, vol. 5, pp. 18429-18438, 2017
[16] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, et al., "IARPA Janus Benchmark-B Face Dataset", 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops, pp. 592-600, 2017
[17] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments", Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, France, Oct 2008
[18] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," 2018 IEEE Int. Conf. on Automatic Face & Gesture Recognition, pp. 67-74, 2018
[19] Y. Wu, Y. Wu, R. Gong, Y. Lv, K. Chen, D. Liang, et al., "Rotation consistent margin loss for efficient low-bit face recognition", 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 6866-6876, 2020
[20] H. Liu, X. Zhu, Z. Lei and S. Z. Li, "AdaptiveFace: Adaptive Margin and Sampling for Face Recognition", 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 11939-11948, 2019
[21] X. Yin, X. Yu, K. Sohn, X. Liu, M. Chandraker, "Feature transfer learning for face recognition with under-represented data", 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019
[22] F. V. Massoli, G. Amato, and F. Falchi, "Cross-resolution learning for Face Recognition", Image and Vision Computing, Vol. 99, 2020
[23] M. Singh, S. Nagpal, M. Vatsa, R. Singh and A. Majumdar, "Identity Aware Synthesis for Cross Resolution Face Recognition", IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2018
[24] S. Kolouri and G. K. Rohde, "Transport-based single frame super resolution of very low resolution face images", 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 4876-4884, 2015
[25] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On Low-Resolution Face Recognition in the Wild: Comparisons and New Techniques", Trans. Info. For. Sec., vol. 14-8, August 2019
[26] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey", ACM Comput. Surv. 46-2, art. 29, Nov. 2013
[27] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey", Neurocomputing, Volume 381, pp. 61-88, 2020
[28] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric", 2017 IEEE Int. Conf. on Image Processing, pp. 3645-3649, 2017
[29] Multiple Object Tracking Benchmark. https://motchallenge.net
[30] D. Mykheievskyi, D. Borysenko, and V. Porokhonskyy, "Learning Local Feature Descriptors for Multiple Object Tracking", 2020 Asian Conf. on Computer Vision, Springer LNCS, vol 12623, 2020
[31] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking Object as Points", 16th Europ. Conf. on Computer Vision, Springer LNCS, vol 12349, 2020
[32] Y. Zhang, D. Zhou, S. Chen, S. Gao and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network", IEEE Conf. on Computer Vision and Pattern Recognition, 2016
[33] L. Boominathan, S. Kruthiventi, and R. Venkatesh Babu, "CrowdNet: A Deep Convolutional Network for Dense Crowd Counting", The 24th ACM Int. Conf. on Multimedia (MM '16), New York, USA, pp. 640–644, 2016
[34] X. Liu, J. van de Weijer and A. D. Bagdanov, "Leveraging Unlabeled Data for Crowd Counting by Learning to Rank", 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 7661-7669, 2018
[35] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, "Almost Unsupervised Learning for Dense Crowd Counting", 2019 AAAI Conf. on Artificial Intelligence, 33(01), pp. 8868-8875, 2019
[36] H. Xiong, H. Lu, C. Liu, L. Liu, Z. Cao and C. Shen, "From Open Set to Closed Set: Counting Objects by Spatial Divide-and-Conquer", 2019 IEEE/CVF Int. Conf. on Computer Vision, pp. 8361-8370, 2019
[37] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu and X. Yang, "Crowd Counting via Adversarial Cross-Scale Consistency Pursuit", 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 5245-5254, 2018
[38] D. Lian, J. Li, J. Zheng, W. Luo and S. Gao, "Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization", IEEE Conf. on Computer Vision and Pattern Recognition, 2019
[39] D. Sam, S. Peri, M. Sundararaman, A. Kamath and R. Babu, "Locate, Size, and Count: Accurately Resolving People in Dense Crowds via Detection", IEEE Tran. on Pattern Analysis & Machine Intelligence, vol. 43, no. 08, pp. 2739-2751, 2021
[40] L. Yuting, W. Zheng, S. Miaojing, S. Shin'ichi, Z. Qijun, and Y. Hongyu, "Towards Unsupervised Crowd Counting via Regression-Detection Bi-knowledge Transfer", 28th ACM Int. Conf. on Multimedia, 2020
[41] Ugur Toreyin, Enis Cetin,"Sample Fire and Smoke Video Clips", http://signal.ee.bilkent.edu.tr/VisiFire/Demo/SampleClips.html
[42] Firesense, https://cordis.europa.eu/project/id/244088