Review on Common Techniques for Urban Environment Video Analytics

Henry O. Velesaca[1], Patricia L. Suárez[1], Angel Sappa[1,2], Dario Carpio[1], Rafael E. Rivadeneira[1], Angel Sanchez [3]

[1]Escuela Superior Politécnica del Litoral, ESPOL,
Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863

[2]Computer Vision Center, Campus UAB
08193 Bellaterra, Barcelona, Spain

[3]E.T.S. Ingeniería Informática, Universidad Rey Juan Carlos
28933 Móstoles, Madrid, Spain

{hvelesac,plsuarez,asappa,dncarpio,rrivaden}@espol.edu.ec, asappa@cvc.uab.es, angel.sanchez@urjc.es

Abstract. *This work compiles the different computer vision-based approaches from the state of the art intended for video analytics in urban environments. The manuscript groups the different approaches according to typical modules present in video analytics including image preprocessing, object detection, classification and tracking. The manuscript is not meant to be an exhaustive review of state-of-the-art approaches but just a list of common techniques proposed to address recurring problems in this field.*

1. Introduction

In recent years, the use of video surveillance applications has increased, either due to the need for security in urban areas or due to technological advances in the area of smarts cameras together with the considerable reduction of their prices. All these facts have allowed the resurgence of smart video surveillance systems together with specialized applications that facilitate the control of many daily operations present in urban environments. Examples of these applications are the face and action recognition, people counting in crowded scenarios, traffic monitoring, license plates recognition, person re-identification, intelligent traffic light management, among others. The market for smart security devices in general and the development of mobile applications is expected to increase from USD 45.5 billion in 2020 to USD 74.6 billion by 2025[1].

With the rise of IP cameras and intelligent sensors, it is possible to collect a large amount of information in real-time that is monitored by specialized personnel, therefore, the analysis of the scenes is susceptible to errors in human perception when monitoring the information and the consequent incorrect decision making, in addition to the loss of time and money. With the specialized video surveillance algorithms, it is possible to analyze the videos in real-time without susceptibility to human error, with which it is possible to react in time, continuously, 24/7, and with superior effectiveness for making the right decisions. Additionally, all the analyzed information can be summarized, stored, and generated the corresponding statistics for historical analysis and continuous improvement of the processes.

[1]https://www.statista.com/statistics/864838/video-surveillance-market-size-worldwide/

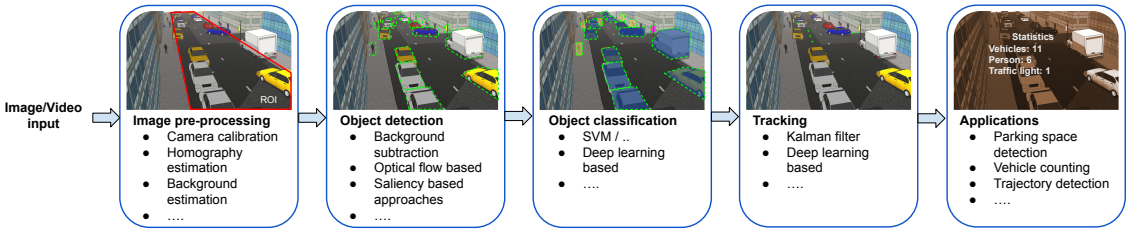

Figure 1. Pipeline with the schematic video analytic modules.

Deep learning techniques are the evolution of machine learning, with them a range of new products have been implemented that address urban video analytics from innovative technological perspectives. In the area of computer vision research, a wide variety of techniques have been developed, some of them focused on understanding multiple videos/scenes. Urban environments today focus on the development of solutions for the detection and tracking of objects, reidentification of people, face recognition, actions or movements made by people, among others. Real-time video analytics applications based on machine learning are highly required in urban environments due to the versatility they can provide in terms of security and information control through video analytics. There are several state-of-the-art approaches depending on the problem to be solved, but most of them share modules such as image pre-processing, including camera calibration, image enhancement, object detection, tracking, and recognition, among others. To guide the development of this work, a pipeline has been proposed with the most representative modules in the state-of-the-art literature to establish order and a common framework when describing the techniques. This paper is organized as follows. First, Section 2 reviews common image preprocessing techniques generally used in this framework. Then, Section 3 summarizes object detection approaches, while Section 4 focuses on the latest classification techniques. Tracking approaches are reviewed in Section 5. Video analytics applications of urban environments are presented in Section 6. Finally, discussions of the challenges and opportunities of the reviewed approaches are given in Section 7.

2. Image Preprocessing

Image preprocessing plays an important role in most computer vision applications. In general, it affects the final results and is responsible for computing more accurate solutions. This section reviews only the most commonly used image processing techniques in video analytics of urban environments. It is not meant to be an exhaustive review but just a list of common approaches proposed to address recurring issues such as camera calibration, background subtraction, and image enhancement.

2.1. Camera Calibration

The precise mapping of 3D points in the image plane, as well as the definition of the region to be analyzed in the image or the relative distance between the objects, is done through the calibration of the camera, which is considered one of the most common approaches in image processing. This calibration process is usually one of the first processes required in setting up the camera using a calibration pattern. It must be considered that the estimation of 2D-3D correspondences that affect real-world applications can be a real limitation. Therefore, approaches that focus on the geometry of the images can be used to perform the calibration, rather than using conventional size-specific standards. An example of

these techniques is the use of lines and vanishing points to perform the camera calibration (e.g., [Caprile and Torre 1990], [Zhang et al. 2015]). With the estimated intrinsic and extrinsic parameters of the camera, homography can be used to reproject pixels from a given image region to a real-world 3D scene. For example, Noh et al. [Noh et al. 2020], propose an inverse perspective mapping approach to compute a top view of crosswalk scenarios. This top view is used to predict possible risk events for pedestrians in the crosswalk environment. The system estimates the trajectories of pedestrians and vehicles once they are detected; then, a decision tree is used to analyze whether or not there is a risk for the pedestrian at the intersection of his path with that of the vehicle. Similarly, Grassi et al. [Grassi et al. 2017] present another camera calibration process using information from the inverse perspective mapping. Calibration information is then used for finding free parking spaces in urban settings, together with GPS location coordinates obtained from smartphones. Finally, a more complex method has been recently presented by Liu et al. [Liu et al. 2020], where speed estimation, vehicle counting, and vehicle classification based on inverse perspective mapping are also proposed; it uses projective geometry together with a deep learning network for vehicle classification.

2.2. Background Subtraction

Another image processing approach, generally applied to static cameras, is the background estimation. Background information is used to subtract from a given frame and easily detect foreground objects, usually moving objects. Although widely used, these approaches have a major limitation, which is the definition of the threshold, which can be a fixed value or a dynamically updated value [Gupte et al. 2002]. Within the great variety of background estimation techniques, one of the simplest ones is the one that calculates the temporal median [Graszka 2014], or the temporal average of consecutive frames [Lee and Hedley 2002]. These simple and intuitive approaches were widely used in the decade of the nineties for traffic surveillance. Unfortunately, this approach is not robust since it does not adapt to multiple scenarios, lighting changes, vibrations, or dynamic backgrounds. More complex background estimation models have been proposed in the literature to overcome these limitations; one of these elaborated models is based on the usage of Mixture of Gaussians (MOG) [Stauffer and Grimson 1999]. The characteristic of these types of models is that each pixel is computed as the combination of two or more Guassianas with online updates. These models support lighting changes and dynamic backgrounds but have a high computational cost compared to simple background estimation approaches. With the rise of deep learning different background estimation approaches have been proposed. For instance, Lim et al. [Lim et al. 2017] propose an encoder-decoder-based convolutional network scheme for segmenting moving objects; modeling the image background by extracting the image foreground information as a feature vector to obtain the segmentation map. This class of models is now considered the gold standard for performance improvements when compared to conventional approaches. Top background subtraction methods evaluated in public benchmarks (e.g., CDnet 2012, CDnet 2014) are deep learning-based approaches. An extensive review of the state of the art of deep learning network-based fundraising methods is presented in Bouwmans et al. [Bouwmans et al. 2019], which collects over four hundred references.

2.3. Image and Content Enhancement

Finally, image enhancement techniques can be also used to improve the quality of captured images in real-time to achieve better results when processed. Some examples of image enhancement for further video analysis are as follows; Qu et al. [Qu et al. 2018] proposes a method to reduce the effects of lighting changes by improving contrast and enhancing the colors of objects present in the image. This approach is applied to detect pedestrians in different scenarios, improving the accuracy of the results. Similarly, another approach that proposes to improve image quality, specifically for night scenarios, is presented in Shi et al. [Shi et al. 2018]. The method is based on the use of dark and light channels with a single image to improve the lack of lighting and achieve images with their content well exposed. Another concern on images in the public environment is privacy protection. In this sense, content manipulation methods are tackling privacy protection issues. With the rise of public camera deployment by all sectors in urban areas with vision systems, privacy and security issues have arisen. Therefore, information protection strategies must be included alongside new ubiquitous implementation approaches. An example of this strategy is the one implemented in Google Street View [Frome et al. 2009], where faces and licenses plates are blurred to preserve the identity of people or vehicles captured in the collected images. Another example of approaches applying face removal to streaming video is proposed in Tu et al. [Tu et al. 2021], where videos can be publicly distributed without compromising data privacy and further processed.

3. Detection

According to the pipeline presented in Fig. 1, after preprocessing the given images, the next process is the detection of objects present in the images. Several approaches have been proposed in the field of object detection. We start with Avola et al proposal, where a method is presented to detect suspicious security behaviors [Avola et al. 2017] in urban or private environments. This technique uses a histogram equalization as well as the RGB local binary pattern operator on high-resolution low-altitude images collected by unmanned aerial vehicles. Also working on high-resolution, Chen et al, presents an object detection method in gigapixel videos to obtain local and global information simultaneously. The approach proposes to find real-time global location by using region descriptors, while local information is obtained using local descriptors. Continuing with detection techniques, in Gautamin et al. [Gautam and Thangavel 2019] a face detection invariant to illumination changes, orientation, and supporting partial occlusions is proposed. For training, a data set of previously registered faces is used, calculating their weight as search criteria. Once the model has been trained, in the validation process, the weights are compared to determine if the faces are coincident or not. The method is implemented with the Haar cascade classifier combined with a skin detector.

Another technique for object detection using a convolutional network is presented in Shi et al. [Shi et al. 2022]; the authors propose to use the geometric information of stereoscopic images to adaptively perform object detection. Results obtained with this technique are quick and accurate enough. Similarly, Bochkovskiy et al. [Bochkovskiy et al. 2020] propose an efficient object detection model in real-world environments, using an architecture that focuses on highlighting the receptive fields of each model layer by applying data augmentation through a 4-image mosaic. The COCO and IMAGENET datasets have been used for the experiments, obtaining better results than its

predecessor, YOLOv3. Instead, in Wei et al. [Wei et al. 2018] an approach to detect and classify people is proposed. Each image is reduced four times and colors are converted to luminance only to improve the detection process. Then, the difference between consecutive images is applied to detect moving areas to identify people, which will be the input of a classification model based on convolutional networks.

On the other hand, an approach to analyze traffic videos in real-time is proposed by Cao et al. [Cao et al. 2020]. It is based on a three-dimensional reconstruction process to obtain the most representative geometric features, which are then used for vehicle detection. Another similar technique is presented in Aslani et al. [Aslani and Mahdavi-Nasab 2013], where real-time detection of moving objects from aerial images is proposed. This approach applies optical flow and a filtering step to remove non-relevant information; a morphological filter is applied to the resulting information to extract the most significant characteristics of the regions of interest and rule out possible occlusions. Another object detection technique, that works in real-time traffic scenarios, to detect objects such as buses, motorcycles, and vehicles is presented by Zhang et al. [Zhang et al. 2018]. To validate the method, three data sets have been used: MS COCO, PASCAL VOC07, KITTI.

In the computer vision research area, object detection is one of the most widely used approaches, generally present in the basis of different applications. This extensive use has generated a large number of specialized applications that combine object detection with other computer vision techniques to solve specific problems such as object classification or segmentation. Likewise, it can be evidenced according to the quantitative and qualitative results that the approaches based on CNN, exceed in precision the classic approaches for the detection of objects. These CNN models facilitate the automatic extraction of the most relevant features and are robust to different scenarios of urban environments. Some solutions of this kind already exist on the market and are useful for building different solutions with them.

4. Classification

Following the proposed pipeline, once the objects in the given images have been detected, the classification process proceeds by analyzing the detected regions of interest to find a pattern that allows their classification. For many years, the object classification process has been a topic of research, which remains active; Varieties of approaches based on machine learning techniques have been proposed. This section covers both classical techniques and those based on deep learning. Beginning with Silva et al. [Silva et al. 2018] where the authors propose an approach to detect and classify motorcycle helmets using a multilayer perceptron network. Similarly, in Cao et al. [Cao et al. 2011], another technique for urban traffic surveillance is presented, where a specialized and robust algorithm is proposed to detect and classify vehicles in environments with varying lighting and with complexity and diversity of scenarios. Continuing with traffic management techniques, in Makhmutova et al. [Makhmutova et al. 2020], a technique based on optical flow using the public network of the urban traffic control system is proposed. With all the collected information it is possible to build models capable of classifying, for example, the types of cars and determining their location, and counting them according to their location in real-time.

Regarding approaches for classifying objects in videos, it could be mentioned the work of Hamida et al. [Hamida et al. 2014]. The authors present a classification

technique, which uses an architecture that can scale according to the client's requirements. This scheme is invariant to spatial changes and focuses on extracting information from the region of interest. Similarly, a novel technique to apply intelligent filtering is proposed in Canel et al. [Canel et al. 2019]. This technique is called FilterForward, which uses a micro classifier that only captures the relevant frames of urban traffic videos to detect coincident events. Another similar filtering technique is presented in Li et al. [Li et al. 2020], which uses thresholds adaptable to variations in the correlation between the types of relevant characteristics during a period of time of the video; it groups the analyzed vehicles in real-time and can classify them. A similar technique to classify vehicles of any type is presented in Buch et al. [Buch et al. 2009]; this approach uses 3D models of the silhouette of each type of vehicle present in the video. Another approach to classifying objects in video surveillance is presented in Xu et al. [Xu et al. 2018]. For the selection of the most relevant features, the linear support vector machine and the oriented gradient histogram are used.

Recently, techniques based on convolutional networks have achieved promising results, which has allowed them to be placed above techniques based on classical approaches. Following this criterion, in Vishnu et al. [Vishnu et al. 2017] the authors introduce a method to detect offending motorcyclists for not wearing a safety helmet. The model uses a convolutional network to perform the classification and recognition of offenders. Another technique that allows predicting possible behavior is presented in Kumar et al. [Kumar 2020]. It is a simple approach capable of processing large amounts of traffic information to estimate the possible behavior and establish a probable forecast of the traffic flow using a convolutional neural network.

5. Tracking

The last stage presented is the object tracking task that uses the information obtained in the previous stages of the pipeline for urban video analytics, this stage is necessary to track the object in the image and detect it correctly to avoid duplication. One of the most used techniques in the object tracking task is the Kalman Filter [Kalman 1960], which is based on the use of state-space techniques and recursive algorithms. On the other hand, the work presented by [Dyckmanns et al. 2011] using Extended Kalman Filter shows an approach to object tracking at urban intersections for detection of incorrect maneuvers by incorporating prior knowledge into the filtering process. Another strategy used for object tracking is the Particle Filter, which estimates the state of a system that changes over time. Rehman et al. [Ata-Ur-Rehman et al. 2021] presents a particle filtering-based anomaly detection framework for online video surveillance, in which they use features of size, location, and motion to develop predictive models and estimate the future probability of activities.

Other works, such as those presented by Jodoin et al. [Jodoin et al. 2014] use techniques based on binary descriptors and background subtraction, which allows road users to be tracked regardless of their size and type. Another recent technique in the object tracking domain is Simple Online and Real-Time Tracking (SORT) [Bewley et al. 2016]. SORT compares the positions of objects detected in the current frame with guesses for the positions of objects detected in previous frames. On the other hand, Velesaca et al. [Velesaca et al. 2020] considers the use of SORT to generate general statistics of the different objects in urban environments.

Contrary to the traditional techniques explained above, deep learning-based techniques have been used in the last decade to significantly increase object tracking performance. Del Pino et al. [del Pino et al. 2017] propose the use of convolutional neural networks to estimate the real position and speeds of detected vehicles. On the other hand, Liu et al. [Liu et al. 2016] propose a deep learning-based progressive vehicle re-identification approach, which uses a convolutional neural network to extract appearance attributes as the first filter, and a Siamese neural network-based license plate verification as the second filter.

All approaches presented above focus their efforts on tracking multiple objects in 2D, that is, in the image plane. Contrary to these approaches, 3D tracking could provide more precise information about the position, size estimation, and effective occlusion management. An example of this approach is the work presented by Nguyen et al. [Nguyen et al. 2011], which uses a stereo camera system to get a cloud of 3D points and map an occupation grid using an inverse sensor model to track different objects within the range of the sensors. However, despite all these benefits, 3D tracking is still developing to solve camera calibration, pose estimation, and scene layout problems.

6. Applications

The application of video analytics in urban environments has received increasing attention in the past few years. This is because video analytics can be used to improve the understanding of the dynamics of urban scenes. According to the modules reviewed above, different video content analysis applications can be applied to process urban environment videos in real-time. This section provides a brief yet concise survey of state-of-the-art applications.

6.1. Traffic Scenarios

License plate recognition is largely used worldwide for vehicle identification, access control, and security, generally in controlled scenarios and stationary vehicles. It is commonly referred to as *automatic number-plate recognition* (ANPR) or *automatic license-plate recognition* (ALPR). This process consists in detecting the license plate in a video frame and then recognizing the number plate of vehicles in real-time videos. Finally, the recognized information is used to check if the vehicle is registered or licensed or even grants access control to private places. The ANPR/ALPR systems have been used in a wide range of applications, places, and institutions, such as shopping malls, police forces, tolls, private places in general, car parking management, and other data collection reasons (e.g., [Liu et al. 2016], [Arafat et al. 2020], [Zou et al. 2021]). Another application is the identification of free parking space, in general, using low-cost IoT devices. These techniques aim to detect free parking slots in real-time, count the number of places for parking management systems, and provide this information to the drivers. Furthermore, in [Liu et al. 2020], vehicle counting, speed estimation, and classification is proposed, which combines object detection and multiple object tracking to count and classify vehicles, pedestrians, and cyclist from video captured by a single camera. Finally, another application for traffic control is the vehicle classification (e.g., cars, buses, vans, motorcycles, bikes) detected in a scene [Buch et al. 2009]; where counting vehicles by lane and estimating vehicle length and speed in real-world units is possible.

6.2. Pedestrian-Oriented Applications

Pedestrian Detection and Tracking are another classical application in urban video analytics environments that has attracted much attention in recent years due to the increase in computational power of intelligent systems. On top of pedestrian detection and tracking, other approaches can be developed, for instance, counting people in specific places or human behavior analysis (e.g., how long they stay in a particular area, where they go, etc.) in intelligent video surveillance systems (e.g., [Gaddigoudar et al. 2017], [Praveenkumar et al. 2022]). Still, these approaches cannot deal with challenging crowded places (occlusions or low-resolution images) or false positives detections; despite these limitations, some applications can get good performance [Liu et al. 2021]. More recently, due to the COVID-19 outbreak, some approaches for pedestrian detection and tracking for social distance have been proposed; for instance, [Ahmed et al. 2021] proposes to use an overhead perspective. This method can estimate social distance violations between people using an approximation of physical distance, making use of the fact that, in general, people walk in a more or less systematic way, and they tend to move in the same direction as the rest of the people in the scene. Another challenging application is the estimation of pedestrian actions on streets [Ridel et al. 2018]. It can be applied to safety issues in smart cars. The estimation of pedestrian actions can provide information such as the intention of the pedestrian (e.g., to cross the street, to stop at a traffic light, etc.), which can be used by smart cars to make decisions about how to behave. Face detection (e.g., [Tu et al. 2021], [Gautam and Thangavel 2019]) is one of the most common applications of video analytics in urban environments. It can be used to identify people who are walking or driving in the area. This information can be used to improve security by tracking specific individuals' movements or improving traffic flow by identifying choke points. Furthermore, face detection can be used to count the number of people in an area, which can be used to estimate the number of people who are likely to visit an area at a given time.

7. Conclusions

Urban growth of large cities makes it very important to have innovative applications for the efficient management of different services. This work summarizes some of the computer vision-based approaches needed to solve problems such as people counting in crowded scenarios, vehicle classification, or detection of free parking spaces. Every year the technology is improving and consequently, the computing power has increased considerably, this has allowed a large number of these approaches to be implemented in real-time or even integrated into low-cost hardware. This work has been organized through a common framework that allows for group contributions in the literature according to the different modules. The review begins with the pre-processing techniques of urban environment images. The object detection and classification approaches are then summarized highlighting several recent techniques, both tasks are merged into a single module. The last module is the tracking of objects once they have been detected and classified. This component is very important because it facilitates decision-making based on the behavior of the object. Initially, tracking objects were performed by Kalman filters, but recently deep learning-based approaches have given the best results. As has been indicated in several sections of the paper, this work is not intended to be a complete list of state-of-the-art approaches for each module of the proposed pipeline, but rather a review of some of the different recent contributions. Regarding traffic scenarios and pedestrian-oriented applications, it

could be mentioned that more and more video analytics applications will be offered in the context of smart cities in the next few years. This growth will happen due to the expansion of camera networks, which is a ubiquitous technology in any urban scenario, the increase in computational capabilities, and the deep-learning strategy generally used to implement the different solutions. These three elements are inspiring the development of novel applications; some of them were unthinkable just a few years ago. Summarizing, it could be stated that solutions based on deep learning are today the ones that present better results than traditional techniques for any of the modules in the pipeline. Nowadays, several opportunities are being generated to be explored from the video analytics of urban environments depending on the complexity of the solution, low-cost hardware embedded applications can be considered.

8. Acknowledgement

This work has been partially supported by the ESPOL projects TICs4CI (FIEC-16-2018) and PhysicalDistancing (CIDIS-56-2020); and the "CERCA Programme/Generalitat de Catalunya". The authors acknowledge the support of CYTED Network: "Ibero-American Thematic Network on ICT Applications for Smart Cities" (REF-518RT0559).

References

Ahmed, I., Ahmad, M., Rodrigues, J. J., Jeon, G., and Din, S. (2021). A deep learning-based social distance monitoring framework for covid-19. *Sustainable Cities and Society*, 65:1–12.

Arafat, M. Y., Khairuddin, A. S. M., and Paramesran, R. (2020). Connected component analysis integrated edge based technique for automatic vehicular license plate recognition framework. *Intelligent Transport Systems*, 14(7):712–723.

Aslani, S. and Mahdavi-Nasab, H. (2013). Optical flow based moving object detection and tracking for traffic surveillance. *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, 7(9):1252–1256.

Ata-Ur-Rehman, Tariq, S., Farooq, H., Jaleel, A., and Wasif, S. M. (2021). Anomaly detection with particle filtering for online video surveillance. *IEEE Access*, 9:19457–19468.

Avola, D., Foresti, G. L., Martinel, N., Micheloni, C., Pannone, D., and Piciarelli, C. (2017). Aerial video surveillance system for small-scale uav environment monitoring. In *14th International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6. IEEE.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In *International Conference on Image Processing*, pages 3464–3468. IEEE.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection.

Bouwmans, T., Javed, S., Sultana, M., and Jung, S. K. (2019). Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. *Neural Networks*, 117:8–66.

Buch, N., Cracknell, M., Orwell, J., and Velastin, S. A. (2009). Vehicle localisation and classification in urban cctv streams. *ITS World Congress*, pages 1–8.

Canel, C., Kim, T., Zhou, G., Li, C., Lim, H., Andersen, D. G., Kaminsky, M., and Dulloor, S. R. (2019). Scaling video analytics on constrained edge nodes. *arXiv preprint arXiv:1905.13536*.

Cao, M., Zheng, L., Jia, W., and Liu, X. (2020). Joint 3d reconstruction and object tracking for traffic video analysis under iov environment. *Transactions on Intelligent Transportation Systems*, 22(6):3577–3591.

Cao, X., Wu, C., Lan, J., Yan, P., and Li, X. (2011). Vehicle detection and motion analysis in low-altitude airborne video under urban environment. *Transactions on Circuits and Systems for Video Technology*, 21(10):1522–1533.

Caprile, B. and Torre, V. (1990). Using vanishing points for camera calibration. *International Journal of Computer Vision*, 4(2):127–139.

del Pino, I., Vaquero, V., Masini, B., Sola, J., Moreno-Noguer, F., Sanfeliu, A., and Andrade-Cetto, J. (2017). Low resolution lidar-based multi-object tracking for driving applications. In *Iberian Robotics Conference*, pages 287–298. Springer.

Dyckmanns, H., Matthaei, R., Maurer, M., Lichte, B., Effertz, J., and Stüker, D. (2011). Object tracking in urban intersections based on active use of a priori knowledge: Active interacting multi model filter. In *Intelligent Vehicles Symposium*, pages 625–630. IEEE.

Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., and Vincent, L. (2009). Large-scale privacy protection in google street view. In *12th International Conference on Computer Vision*, pages 2373–2380. IEEE.

Gaddigoudar, P. K., Balihalli, T. R., Ijantkar, S. S., Iyer, N. C., and Maralappanavar, S. (2017). Pedestrian detection and tracking using particle filtering. In *International Conference on Computing, Communication and Automation*, pages 110–115.

Gautam, K. and Thangavel, S. K. (2019). Video analytics-based intelligent surveillance system for smart buildings. *Soft Computing*, 23(8):2813–2837.

Grassi, G., Jamieson, K., Bahl, P., and Pau, G. (2017). Parkmaster: An in-vehicle, edge-based video analytics service for detecting open parking spaces in urban environments. In *Proceedings of Symposium on Edge Computing*, pages 1–14.

Graszka, P. (2014). Median mixture model for background–foreground segmentation in video sequences. *22nd International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision in co-operation with EUROGRAPHICS Association*, pages 103–110.

Gupte, S., Masoud, O., Martin, R. F., and Papanikolopoulos, N. P. (2002). Detection and classification of vehicles. *Transactions on Intelligent Transportation Systems*, 3(1):37–47.

Hamida, A. B., Koubaa, M., Amar, C. B., and Nicolas, H. (2014). Toward scalable application-oriented video surveillance systems. In *Science and Information Conference*, pages 384–388. IEEE.

Jodoin, J.-P., Bilodeau, G.-A., and Saunier, N. (2014). Urban tracker: Multiple object tracking in urban mixed traffic. In *Winter Conference on Applications of Computer Vision*, pages 885–892. IEEE.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

Kumar, T. S. (2020). Video based traffic forecasting using convolution neural network model and transfer learning techniques. *Journal of Innovative Image Processing*, 2(03):128–134.

Lee, B. and Hedley, M. (2002). Background estimation for video surveillance. *Image Vision Computing New Zealand*, pages 315–320.

Li, Y., Padmanabhan, A., Zhao, P., Wang, Y., Xu, G. H., and Netravali, R. (2020). Reducto: On-camera filtering for resource-efficient real-time video analytics. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, pages 359–376.

Lim, K., Jang, W.-D., and Kim, C.-S. (2017). Background subtraction using encoder-decoder structured convolutional neural network. In *14th International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6. IEEE.

Liu, C., Huynh, D. Q., Sun, Y., Reynolds, M., and Atkinson, S. (2020). A vision-based pipeline for vehicle counting, speed estimation, and classification. *Transactions on Intelligent Transportation Systems*.

Liu, X., Liu, W., Mei, T., and Ma, H. (2016). A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European conference on computer vision*, pages 869–884. Springer.

Liu, X., Sang, J., Wu, W., Liu, K., Liu, Q., and Xia, X. (2021). Density-aware and background-aware network for crowd counting via multi-task learning. *Pattern Recognition Letters*, 150:221–227.

Makhmutova, A., Anikin, I. V., and Dagaeva, M. (2020). Object tracking method for videomonitoring in intelligent transport systems. In *International Russian Automation Conference*, pages 535–540. IEEE.

Nguyen, T.-N., Michaelis, B., Al-Hamadi, A., Tornow, M., and Meinecke, M.-M. (2011). Stereo-camera-based urban environment perception using occupancy grid and object tracking. *Transactions on Intelligent Transportation Systems*, 13(1):154–165.

Noh, B., No, W., Lee, J., and Lee, D. (2020). Vision-based potential pedestrian risk analysis on unsignalized crosswalk using data mining techniques. *Applied Sciences*, 10(3):1–21.

Praveenkumar, S., Patil, P., and Hiremath, P. (2022). Real-time multi-object tracking of pedestrians in a video using convolution neural network and Deep SORT. In *ICT Systems and Sustainability*, pages 725–736. Springer.

Qu, H., Yuan, T., Sheng, Z., and Zhang, Y. (2018). A pedestrian detection method based on YOLOv3 model and image enhanced by retinex. In *11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pages 1–5. IEEE.

Ridel, D., Rehder, E., Lauer, M., Stiller, C., and Wolf, D. (2018). A literature review on the prediction of pedestrian behavior in urban scenarios. In *21st International Conference on Intelligent Transportation Systems*, pages 3105–3112. IEEE.

Shi, Y., Guo, Y., Mi, Z., and Li, X. (2022). Stereo CenterNet-based 3D object detection for autonomous driving. *Neurocomputing*, 471:219–229.

Shi, Z., Guo, B., Zhao, M., Zhang, C., et al. (2018). Nighttime low illumination image enhancement with single image using bright/dark channel prior. *Journal on Image and Video Processing*, 2018(1):1–15.

Silva, R. R., Aires, K. R., and Veras, R. d. (2018). Detection of helmets on motorcyclists. *Multimedia Tools and Applications*, 77(5):5659–5683.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252. IEEE.

Tu, N. A., Wong, K.-S., Demirci, M. F., Lee, Y.-K., et al. (2021). Toward efficient and intelligent video analytics with visual privacy protection for large-scale surveillance. *The Journal of Supercomputing*, pages 1–31.

Velesaca, H. O., Araujo, S., Suárez, P. L., Sánchez, A., and Sappa, A. D. (2020). Off-the-shelf based system for urban environment video analytics. In *International Conference on Systems, Signals and Image Processing*, pages 459–464.

Vishnu, C., Singh, D., Mohan, C. K., and Babu, S. (2017). Detection of motorcyclists without helmet in videos using convolutional neural network. In *International Joint Conference on Neural Networks*, pages 3036–3041. IEEE.

Wei, H., Laszewski, M., and Kehtarnavaz, N. (2018). Deep learning-based person detection and classification for far field video surveillance. In *13th Dallas Circuits and Systems Conference*, pages 1–4. IEEE.

Xu, R., Nikouei, S. Y., Chen, Y., Polunchenko, A., Song, S., Deng, C., and Faughnan, T. R. (2018). Real-time human objects tracking for smart surveillance at the edge. In *International Conference on Communications*, pages 1–6. IEEE.

Zhang, H., Wang, K., Tian, Y., Gou, C., and Wang, F.-Y. (2018). MFR-CNN: Incorporating multi-scale features and global information for traffic object detection. *Transactions on Vehicular Technology*, 67(9):8019–8030.

Zhang, M., Yao, J., Xia, M., Li, K., Zhang, Y., and Liu, Y. (2015). Line-based multi-label energy optimization for fisheye image rectification and calibration. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, pages 4137–4145. IEEE Computer Society.

Zou, Y., Zhang, Y., Yan, J., Jiang, X., Huang, T., Fan, H., and Cui, Z. (2021). License plate detection and recognition based on YOLOv3 and ILPRNET. *Signal, Image and Video Processing*, pages 1–8.