Video Analytics in Urban Environments: Challenges and Approaches

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

Abstract This chapter reviews state-of-the-art approaches generally present in the pipeline of video analytics on urban scenarios. First, a typical pipeline is proposed to cluster approaches in the literature, including image preprocessing, object detection, object classification, and object tracking modules. Then, a review of recent approaches for each module is given. Additionally, applications and datasets generally used for training and evaluating the performance of these approaches are included. This chapter does not pretend to be an exhaustive review from state-of-the-art video analytics in urban environments but rather an illustration of some of the different recent contributions. The chapter concludes by presenting current trends in video analytics in the urban scenario field.

1 Introduction

The reduction in the price of cameras has led to the extension of camera networks to different applications in urban environments. Actually, cameras have become a ubiquitous technology in our everyday life. Although they are mainly used for video surveillance applications, growing in popularity opens new possibilities for smart cities (e.g., from detection of available parking to road maintenance applications). In terms of economy, the global video surveillance market size is expected to grow from USD 45.5 billion in 2020 to USD 74.6 billion by 2025. Increasing concerns about public safety and security, the growing adoption of IP cameras, and the rising demand for wireless sensors are the factors driving the growth of the video surveil-

ESPOL Polytechnic University, FIEC, CIDIS, Guayaquil, Ecuador
e-mail: {hvelesac, plsuarez, dncarpio, rrivaden, asappa}@espol.edu.ec
· E.T.S. Ingeniería Informática, Universidad Rey Juan Carlos, 28933 Móstoles, Madrid, Spain
e-mail: angel.sanchez@urjc.es
· Computer Vision Center, 08193-Bellaterra, Barcelona, Spain
e-mail: asappa@cvc.uab.es

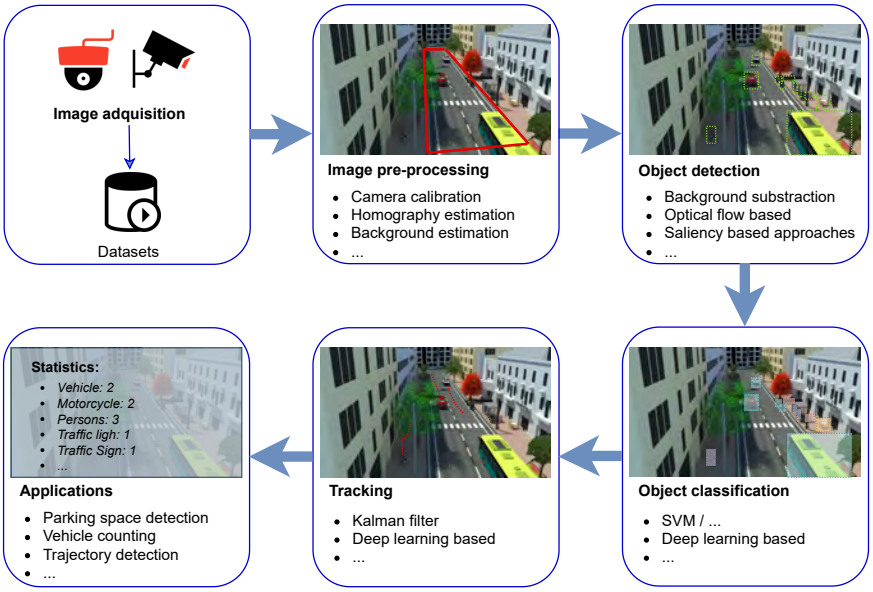

Fig. 1 Pipeline with the modules reviewed in the current work.

lance industry. This huge amount of data is generally used for real-time monitoring by video-surveillance operators or for offline review if something needs to be investigated. All in all, after a user-defined time, which depends on the infrastructure, this information is eliminated from storing systems, missing the opportunity to transform video signals into a powerful source of information. Video analytics engines perform this transformation of signal into information. Video analytics automatically enhances video surveillance systems by performing the tasks of real-time event detection, post-event analysis, and extraction of statistical data while saving human resources costs and increasing the effectiveness of the surveillance system operation.

Developments related to technological advancements like deep learning have increased the feasibility of launching new and innovative products for urban video analytics. Among the vast and rich computer vision research topics in the video/scene understanding domain, the essential needs continue to focus on object detection and tracking, person re-identification and recognition for faces, actions, and license plates recognition, among others. In many cases, the mentioned techniques need to be run in the wild, which means a huge variety of scene conditions, and to achieve a proper and robust solution is not trivial at all. Even though robust object detection and tracking are crucial for any automatic surveillance application, this can be seen as a preliminary step to unlock various use cases.

Video surveillance with machine learning can be a very cost-effective solution for different applications building on top of video analytics of urban environments. Although each one of these applications requires their specific implementations, there are common processing modules that are generally used, for instance, image preprocessing, including camera calibration and image enhancement, object detection and recognition, tracking, to mention a few. Hence, in this chapter, in order to review the state-of-the-art approaches, a module-based architecture is proposed, which clusters the most common approaches in the literature to solve specific tasks (see Fig. 1). This module-based architecture allows the analysis of state-of-the-art approaches and reviews them in a common framework. The current chapter is organized as follows. First, Section 2 reviews common image preprocessing techniques generally used in this framework. Section 3 summarizes object detection approaches, while Section 4 focuses on classification techniques from the state-of-the-art. In Section 5 tracking approaches are reviewed. Video analytics applications of urban environments are presented in Section 6. Although in this chapter there is not a review on image acquisition hardware, common datasets generally used as benchmarks to evaluate different approaches are depicted in Section 7. Finally, discussions on challenges and opportunities from the reviewed approaches are given in Section 8

2 Image Preprocessing

An important stage for obtaining further accurate solutions is the image processing applied to the given input image before performing any computer vision approach. The most common image processing approach is the camera calibration, needed to accurately map 3D points to the image plane and define the region of the image to be analyzed or the relative distance between objects in the scene. In general, camera calibration is performed at the initial setup using a calibration pattern. The need for specific calibration patterns with known sizes for estimating the 2D-3D correspondences is a clear disadvantage for real-world applications. In order to overcome this limitation, there are some approaches that exploit the geometry from the images to perform the calibration, for instance using vanishing point and lines (e.g., [16], [72], [68], [23]). Once extrinsic and intrinsic camera parameters are estimated, homography can be applied to reproject pixels from a given image region to the 3D real-world scenario. For instance, in [55] the authors propose an inverse perspective mapping to obtain a top view of crosswalk regions used for potential pedestrian risky event analysis. The system detects vehicles and pedestrians and estimates their trajectories; then, a decision tree is used to analyze whether there is or not a risk for the pedestrian at the crosswalk (Fig. 2 shows an illustration of the top view computed after camera calibration at a crosswalk). Another camera calibration-based approach, but in this case for inverse perspective mapping from on-board camera systems, is considered in [33], where top view images together with GPS location (smartphone information) are used to identify parking spaces

in urban scenarios. A more elaborated approach has been recently presented in [49], where also an inverse perspective mapping approach is proposed for vehicle counting, speed estimation, and classification. The proposed approach combines a convolutional neural network classifier with projective geometry information to classify vehicles.

On the contrary to previous approaches where bird's-eye views, obtained from inverse perspective mapping, are considered, in [35] just the oblique view is used to obtain the polygonal region of interest from the given image. A local reference frame is placed over this polygonal ROI to estimate the pose of the objects in the scene. The approach allows to count, classify, and determine the speed of vehicles in the annotated ROI. In [48], perspective images are used for identifying free parking spaces in a low-cost IoT device. The system can provide real-time traffic and parking management information for smart cities.

Background estimation is another image processing approach generally used with static cameras. Background models are used to subtract from a given frame and easily detect the foreground objects (moving objects). The main limitation with these approaches is the threshold definition; it can be a constant value or dynamically updated [37]. Different background models have been proposed in the literature, one of the simplest and easy to compute is the temporal averaging of consecutive frames [44] or the temporal median [34]. These methods were widely used in the 1990s for traffic surveillance due to their simplicity. However, they are not robust to challenging scenarios, including fast changes in lighting conditions, camera jitter, or dynamic backgrounds. More elaborated background models have been proposed by using statistical models such as Mixture of Gaussians (MOG) [62]. In these methods, each pixel is modeled as a mixture of two or more Gaussians and online updates. These methods allow modeling scenarios with dynamic backgrounds or with lighting changes. Their main limitation is the computational complexity, resulting in a higher processing time than simpler background subtraction approaches. Recently, deep learning-based approaches have been proposed for the background subtraction problem. For instance, in [47] an encoder-decoder structured CNN has been proposed to segment moving objects from the background. The network models image background while extracting foreground information as a high-level feature vector used to compute a segmentation map. Deep learning-based approaches have shown a considerable performance improvement compared with conventional unsupervised approaches; actually, top background subtraction methods evaluated in public benchmarks (e.g., CDnet 2012, CDnet 2014) are deep learning-based approaches. An extensive study on background subtraction methods based on deep neural networks is presented in [11], where more than four hundred papers are reviewed.

The ubiquitous deployment of vision systems in public areas has increased privacy and security concerns. Hence, data protection strategies need to be appropriately designed and correctly implemented in order to mitigate this problem. An example of a strategy implemented to preserve identity at a scale is the face and license plate blurring process in Google Street View [28]. New intelligent surveillance systems are being designed to face up this concern; for instance, in [63] a privacy-preserving

(a) (b)

Fig. 2 Example of inverse perspective mapping from (a) oblique view; into (b) top view.

video streaming strategy has been proposed. It consists of face detection and removal strategy; hence, videos can be later on distributed for further processing.

Finally, other image processing approaches generally applied in the video analysis of urban environments are related to the image enhancement processes; these image processing approaches are intended to improve the quality of the image in order to reach a better final result. In [57] an image enhancement processing stage is performed to reduce the influence of light changes. The proposed approach enhances the image's contrast and highlights the objects' color. The approach is tested on pedestrian detection in different scenarios, improving the accuracy of the results. Another image enhancement approach, also focusing on lighting problems, more specifically intended for nighttime low illumination scenarios, has been proposed by [61]. The authors propose a dual-channel prior-based method—dark and bright channels—with a single image. The proposed approach corrects the lack of illumination as well as gets well-exposed images. Mapping the given image to another color space is a common image processing approach, generally used to enhance the representation of the image to facilitate further processing. For instance, [74] presents an image enhancement algorithm based on multi-scale Retinex in HSV color model. The authors propose a novel approach that requires a less computational cost to convert from the RGB to HSV and back again to RGB, this mapping to the HSV space is required to correct the V component. The proposed approach has been tested in different outdoor scenarios showing appealing results. Finally, in [21], the authors propose an image processing approach to detect and suppress shadows in a surveillance system. The technique is performed in the HSV color space, which improves the precision of the location of the shadows, favors the extraction of the most relevant information, and facilitates further processing.

3 Detection

Once the information has been preprocessed, the next step, according to the pipeline presented in Fig. 1, is the detection of objects before their classification. Different approaches have been proposed in the literature. The authors in [27], propose a

technique to detect abandoned objects in public places, for which a temporary static objects model is taken as a basis to simulate in an urban environment the complete cycle of people or cars stopped for momentary causes such as traffic lights or signals using a finite state machine. This approach allows detecting the object over a given region directly without applying a classification approach. This model mitigates the problem of detecting false positives in urban environments. In [6] the authors present an approach to detect anomalous security behaviors in public or private environments. The detection technique uses histogram equalization and RGB local binary pattern operator based on images' changes. High-resolution images taken at low altitudes by unmanned aerial vehicles have been used to demonstrate the approach. Another approach to detecting objects is presented in [18], where it is proposed to use gigapixel video; this type of high-resolution image provide local and global information simultaneously to obtain in real-time the location of the object globally by regions and then using descriptors the location of the object is complemented locally.

In [40] the authors propose an approach based on convolutional networks, where a lightweight model has been designed that focuses on increasing the receptive field of the network through the use of a pyramidal combination of the layers of the model. The experiments were favorable to be implemented in real-time applications. In [17], the authors propose a technique for detecting moving objects and subsequent monitoring. The deep learning model proposed by the authors called RBF-FDLNN has been used. This model allows it to extract the necessary characteristics to detect moving objects, allowing their subsequent monitoring efficiently. The datasets used in the experiments are PASCAL and KITTI, which contain objects in urban settings. Another approach to object detection is presented in [60]; the authors propose a convolutional network model using the geometric information of stereo images. This approach does not require anchor boxes or depth estimation, it adaptively performs the detection of objects, achieving speed and precision in the obtained results. In [9], the authors propose an object detection model capable of supporting real-world environments efficiently, for which they have designed an architecture that focuses on enhancing the receptive fields of each layer, also using drop blocks as a regularization method. They have also applied data augmentation through the mosaic of 4 images. The blocks have been designed for punctual attention. The COCO and Imagenet datasets have been used for the experiments, obtaining better statistics than their predecessor, YOLOv3.

In Wei et. al, [67] an approach to detect and classify people is proposed. For the detection step, several operations have been applied to reduce the computational load, for which each image is reduced four times and the colors are converted to only luminance. It applies the difference between consecutive images to detect moving areas and identify people, which are then classified with a convolutional network model. Another approach is presented in [14], where the authors present a method to analyze videos of traffic scenarios. It is based on a three-dimensional reconstruction model to extract geometric features, which is later on used for vehicle detection. In [4] another approach that performs the real-time detection of moving objects from aerial images is presented. It is based on the usage of optical flow and a filtering step

to remove non-relevant information; a morphological filter is applied to extract the most significant features of the regions of interest and discard occlusions.

In Gao et. al [30] a method to detect and count people is presented; this approach is a combination of two techniques since it uses a convolutional network for the extraction of characteristics, in this case, the MASK R-CNN and the linear support vector machine to perform the detection of people. As input for the deep learning model, the regions of people's heads obtained from the images by means of the Adaboost algorithm are used. The authors generate their data set taken from real classroom surveillance scenes for the experiments. Another method that uses convolutional networks is presented in [71], where a method to detect objects (e.g., buses, cars, and motorcycles) in traffic scenarios are presented. The focus is on two optimized components that allow real-time object detection. The first component generates the candidate regions, and the second component detects the objects in the previously selected regions. To improve the performance of the method, global and multi-scale information is incorporated from the processed videos. Three datasets have been used to validate the proposed method: MS COCO, PASCAL VOC07, KITTI.

4 Classification

Once the regions of interest have been detected, they need to be analyzed and classified according to the categories defined in the given problem. For decades, object classification has been an active research topic; several approaches have been proposed based on different machine learning techniques. This section covers both classical and deep-learning-based techniques. In [26], the authors propose an approach to detect and classify motorcycle helmets using a multi-layer perceptron network. Also focusing on the classification problem, in [15], a novel approach is proposed to visualize the trajectory of vehicles in an augmented reality framework by reconstructing objects classified as vehicles in 3D. Another technique for urban traffic surveillance is presented in [15], where the authors propose a robust algorithm to detect and classify vehicles in environments of variation of lighting and complicated scenes, to maintain the most representative global characteristics to accelerate the detection and classification of the vehicles. Low altitude aerial images have been processed to carry out the experiments applying the gradient histogram method. In [52], also tackling the traffic management problem, an optical flow-based approach is proposed by using the cameras of the public traffic control system. The main idea is to classify the types of cars to determine the location and count them later.

Continuing with the classification of objects in videos, a method is proposed in [10], which allows the use of low-resolution images with projective distortion to classify objects present in the far-field. This method makes it possible to classify pedestrians and vehicles invariant to changes in the scenes through contextual functions based on the position and direction of movements of the objects present in the videos. Another classification technique is presented in [38], where the authors pro-

pose an architecture that can scale according to the client's requirement. This scheme does not affect the temporal, spatial, or qualitative aspects and focuses on extracting the information from the region of interest, which is filtered to maintain only the most representative characteristics. In [13], a novel method is presented to perform intelligent filtering. It is referred to as FilterForward, which using micro-classifiers takes only relevant frames from the urban traffic videos, with which coinciding events (i.e., objects of specific colors carried by pedestrians) are detected. Also, an approach to classify vehicles (e.g., buses, cars, motorcycles, vans) is presented in [12]; this method is based on 3D models of the silhouette of each of the vehicle models present in the video. To validate the efficacy of the designed model, the UK reference i-LIDS dataset was used.

On the contrary to previous approaches, in [31] a face detection method that is invariant to changes in illumination, orientation and supports partial occlusions is proposed. The approach is based on training with authorized faces calculating their weight; then, in the validation process, the weights are compared to determine whether or not the faces match. The process of classifying the faces is carried out by applying the Haar cascade classifier combined with the skin detector. Additionally, in [70], an approach is presented to classify in real-time objects present in a surveillance video. Linear support vector machine and the histogram of oriented gradients have been used for the feature extraction. In [46], the authors propose a threshold filtering method adaptable to variations in correlation between the types of features presented over a period of time; this approach is based on clustering, which allows to carry out the real-time analysis of classified vehicles.

In addition to the classical approaches mentioned above, many convolutional neural network-based approaches have been proposed in recent years. In [65] the authors present a technique to detect motorcycle offenders for not wearing a helmet. The model is based on a convolutional network model to perform the classification and recognition of drivers without a helmet. In [43] a simplified approach to process a large amount of traffic data to make the prediction and forecast of its probable behavior are presented. The authors use a convolutional neural network model to predict and forecast urban traffic behavior. In [45] a network, which extends the framework of the faster R-CNN architecture, has been proposed to detect and classify any object presented in each frame of the urban video sequence. Another approach that presents a method for estimating the volume of urban traffic is described in [73]. This technique is based on a deep learning model that works with high-resolution images taken from an unmanned aerial vehicle, which were manually tagged. This model recognizes and classifies vehicle types in trucks, cars, or buses. Similarly, in [66], the authors present an approach based on a convolutional network called NormalNet, which has been designed to extract the characteristics of images using 3D information combined with normal vectors to achieve an efficient classification. Another CNN approach is presented in [53], where the authors implement a customization work of the YOLOv4 model to perform multiple object detection using a low-resolution real-time video, with which nine different types of vehicles are classified. Continuing with deep learning in [3], the authors present a technique to collect information from traffic videos in real-time to estimate the type of vehicle, its

speed, and the respective count. A convolutional network based on regions of interest known as Faster RCNN has been designed for this approach. More recently, in [24], a technique based on transfer learning has been proposed. It obtains the information from previously trained models to be able to classify the vehicles present in the urban traffic videos in real-time. Foreground region detection is applied to group vehicle patterns to achieve vehicle classification.

5 Tracking

The last stage to obtain an accurate solution is the object tracking task using the information obtained in the previous stage of the pipeline. Since the objective of urban video analytics is related to counting, identifying, or re-identifying objects in urban environments (e.g., cars, trucks, buses, motorcycles, bicycles, pedestrians, and other moving objects), tracking the object is needed once it appears in the image and is correctly detected to avoid duplicity. Generally, it is not only necessary to track a single object in the sequence of frames, so Multiple Object Tracking (MOT) is a task that currently plays an essential role in computer vision. MOT task is mainly divided into locating multiple objects, maintaining their identities, and performing their trajectories given a video input.

Different types of probabilistic inference models are used in different works to carry out the MOT task.The most widely used technique is the Kalman Filter [41], which is a recursive predictive filter that is based on the use of state-space techniques and recursive algorithms. To carry out the tracking, [19] uses the centroid of each detected blob using a constant velocity Kalman Filter model. The state of the filter is the centroid location and velocity, and the measurement is an estimate of this entire state. A variation of the original Kalman Filter technique called Extended Kalman Filter (EFK) is used in other works. The authors in [20] merge the measurement obtained from the different sensors (e.g., cameras, radars, and lidars) to track neighboring objects; and the EFK implementation uses the sequential sensor method that treats the observations of individual sensors independently and feeds them sequentially to the EKF estimation process. Also, based on EFK, in [25] the authors present an approach to object tracking in urban intersections, demonstrating that maneuvers can be recognized earlier, and incorrect maneuver detection can be avoided by incorporating prior knowledge into the filtering process. The active multiple model filter approach helps to compensate for unavoidable measurement errors and thus enables robust and stable estimations of the position and velocity of tracked objects. On the other hand, the work presented in [54] uses an Interacting Multiple Model (IMM) [8] because objects such as cars, trucks, bikers, etc., change their behavior so frequently that one motion model alone is not sufficient to track their movements reliably. Structurally, the IMM tracking framework consists of several EKFs with different motion models, and the track estimates are mixed statistically.

Another strategy used for object tracking is the Particle Filter, a method used to estimate the state of a system that changes over time. More specifically, it is

a Monte Carlo method commonly used in computer vision for tracking objects in image sequences. The authors in [5] present an anomaly detection framework based on particle filtering for online video surveillance, where they use characteristics of size, location, and movement to develop prediction models and estimate the future probability of activities in video sequences. On the other hand, [29] uses the particle filter algorithm to track fast-moving human targets; mainly, it performs better with pedestrians with occlusions.

Other works such as the one presented in [39] use classical techniques based on binary descriptors and background subtraction, the combination of both methods allows us to track road users independently of their size and type. Another recent technique in the object tracking domain is Simple Online and Real-Time Tracking (SORT) [7]. This algorithm is based on the Kalman Filter [41] and Hungarian Algorithm [42]. SORT compares the positions of the objects detected in the current frame with assumptions for the positions of the objects detected in previous frames. The authors in [35] present a study that addresses the problem of estimating traffic flows from data from video surveillance cameras; in this work, the Fast R-CNN is used to detect objects, and later on, the SORT is considered to estimate the total number of vehicles in an avenue in real-time. On the other hand, in [64] the usage of SORT is considered to generate general statistics of the different objects in the scene; that the COCO dataset contains (e.g., 91 categories) using Yolo V3 as an object detector. Among the categories for urban environments there are person and vehicle.

Contrary to the traditional techniques explained above, in the last decade, techniques based on deep learning have been used to significantly increase the performance of object tracking in urban environments. The authors in [22] propose the use of convolutional neural networks to estimate the actual position and velocities of the detected vehicles. On the other hand, [50] proposes 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 a second filter. The SORT algorithm presented above was later enhanced to add a deep association metric, and this new approach was named DeepSORT [69]. This model adds visual information extracted by a custom residual CNN. CNN provides a normalized vector with 128 features as output, and the cosine distance between those vectors was added to the affinity scores used in SORT. In the work presented by in [56], DeepSORT is used to track multiple pedestrians in real-time, using motion representation and data association algorithms. The virtual block counting method is implemented for efficient tracking and counting of multiple pedestrian objects, in which the occlusion problem is effectively solved. In other recent works, DeepSORT was used to track vehicles, thus determining the number and flow of cars that travel on a particular avenue [59].

All approaches presented above focus their efforts on tracking multiple objects in 2D, that is, in the image plane. Contrary to this approach, 3D tracking could provide more precise information about the position, size estimation, and effective occlusion management; this approach is potentially more helpful for high-level machine vision tasks. An example of this approach is the work presented in [54], which uses a

Fig. 3 Illustration of parking space detection in a parking lot (10 free space highlighted in red bbox).

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. A wide range of applications can benefit from results of detection, classification, and tracking techniques; just some examples of these applications are: Parking Space Detection, Vehicle or People Counting, Trajectory Detection, Face recognition, Vehicle License Plate Recognition, among others. This section provides a brief yet concise survey of state-of-the-art applications.

For vehicle identification, access control, and security, license plate recognition is largely used worldwide, in general in controlled scenarios and with static vehicles. It is commonly referred to as *automatic number-plate recognition* (ANPR) or *automatic license-plate recognition* (ALPR). This process, as shown in Fig. 4, 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 grant access control to private places. The ANPR/ALPR systems have been used in a wide range of applications, places, and

Fig. 4 Illustration of vehicle license plate recognition process for access control.

institutions, such as shopping malls, police forces, tolls, private places in general, car parking management, and other data collection reasons (e.g., [50], [2], [75]). The main purpose of these systems is to identify vehicles and their registered owners by reading the license plates. The technology is used by police forces to identify stolen cars, by customs to detect illegal immigrants, and by parking companies to issue parking tickets. Another urban environment application based on video analytics is the identification of free parking space (as shown in Fig. 3), in general, using low-cost IoT devices. These techniques aim to detect in real-time free parking slots and count the number of places for parking management systems and provide this information to the drivers. This task is difficult because of the large number of parking spaces and the complexity of the parking environments. License plate detection is also required for privacy protection [28], detecting and blurring the plate number for confidentiality reasons. Furthermore, in [49], vehicle counting, speed estimation, and classification is proposed, which combines object detection and multiple objects 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 [12]; were counting vehicles by lane and estimating vehicle length and speed in real-world units is possible.

Pedestrian detection and tracking are another classical application in urban video analytic 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 (Fig. 5) 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., [29], [56]). Still, a limitation of these approaches is to deal with challenging crowded places (occlusions or low-resolution images) or false positives detections; in spite of these limitations, some applications can get good performance [51]. More recently, due to the COVID-19 outbreak, some approaches for pedestrian detection and tracking for social distance have been proposed; for instance, [1] 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 ordered way, and they tend to move in the same

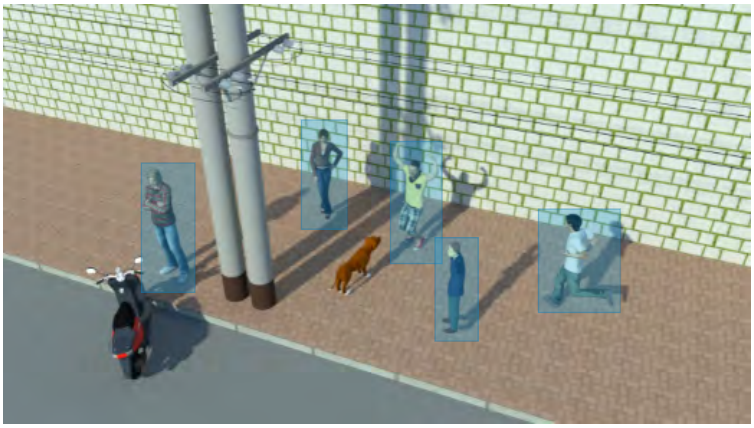

Fig. 5 Illustration of detecting and counting people in the street (5 persons detected).

direction as the rest of the people in the scene. Another challenging application is the estimation of pedestrian actions on streets [58]. It can be applied for safety issues on 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., [63], [31]) 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 to improve 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 Datasets

This section details some of the different datasets used in the state-of-the-art video analytics in urban environments to detect, classify, and track different types of objects (e.g., pedestrians, vehicles, license plates, traffic signs). It does not pretend to be an exhaustive list of datasets but rather an illustration of the different benchmarks available for evaluation and comparisons. One of the most used and popular datasets is the **KITTI**[1] dataset (e.g., [22], [71], [17], [60]) that allow tracking of both people and vehicles. The dataset has been collected by driving a car around a city. It consists of 21 training videos and 29 test ones, with a total of about 19000 frames. Item includes detections obtained using the DPM and RegionLets detectors, as

[1] http://www.cvlibs.net/datasets/kitti/eval_tracking.php

well as stereo and laser information. Another of the widely used datasets is **MS COCO**[2] (Microsoft Common Objects in Context) dataset (e.g., [71], [9]), which is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of 328K images. **PASCAL VOC 2007**[3] dataset (e.g., [71], [17]) is generally used for training image recognition approaches. This dataset includes twenty object classes to find persons, animals, vehicles, and others. The dataset can also be used for image classification and object detection tasks. Another of the popular datasets is **ImageNet**[4], which is an image database (e.g., [9], [24]) organized according to the WordNet hierarchy (currently only the nouns), in which hundreds and thousands of images depict each node of the hierarchy. The project has been instrumental in advancing computer vision and deep learning research. Another dataset used in the literature is **UA-DETRAC**[5] (e.g., [45]), which consists of 100 video sequences captured from real-world traffic scenes (over 140000 frames with rich annotations, including occlusion, weather, vehicle category, truncation, and vehicle bounding boxes) for object detection, object tracking and MOT system.

More specifically for urban environments, **VeRi-776**[6] is a vehicle re-identification dataset (e.g., [50]) that contains 49357 images of 776 vehicles from 20 cameras. The dataset is collected in the real traffic scenario, which is close to the setting of CityFlow. The dataset contains bounding boxes, types, colors, and brands of each vehicle. Another interesting dataset, which involve vehicle images captured across day and night, is the **Vehicle-1M**[7] dataset (e.g., [36]). It uses multiple surveillance cameras installed in cities. There are 936051 images from 55527 vehicles and 400 vehicle models in the dataset. Each image is attached with a vehicle ID label denoting its identity in the real world and a vehicle model label indicating the vehicle's car maker, model, and year. On the contrary to previous datasets **AVSS-2007**[8] dataset (e.g., [27]) is focused on subway environments. It includes three video clips corresponding to low, intermediate, and high activity in a subway scenario. Each clip is about 2 to 3 minutes long, with one staged true drop. Another of the datasets widely used for evaluations and comparisons is **CDnet**[9] dataset, which is an expanded change detection benchmark dataset (e.g., [47]). Identifying changing or moving areas in the field of view of a camera is a fundamental preprocessing step in computer vision and video processing. Example applications include visual surveillance (e.g., people counting, action recognition, anomaly detection, post-event forensics), smart environments (e.g., room occupancy monitoring, fall detection, parking occupancy detection), and video retrieval (e.g., activity localization and tracking). On the other

[2] https://cocodataset.org/#home

[3] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/

[4] https://www.image-net.org/

[5] https://detrac-db.rit.albany.edu/

[6] https://vehiclereid.github.io/VeRi/

[7] http://www.nlpr.ia.ac.cn/iva/homepage/jqwang/Vehicle1M.htm

[8] http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007.html

[9] http://www.changedetection.net/

hand, **PANDA**[10] is the first gigaPixel-level human-centric video dataset (e.g., [18]), for large-scale, long-term, and multi-object visual analysis. The videos in PANDA were captured by a gigapixel camera and cover real-world large-scale scenes with both wide field-of-view and high-resolution details (gigapixel-level/frame). The scenes may contain 4K headcounts with over 100× scale variation.

As mentioned above, this is just an illustration of many benchmarks available for the research community. Table 1 presents a summary of the approaches found in the literature for each of the modules in Fig. 1; furthermore, this table includes datasets and research papers that use them.

8 Conclusions

Novel solutions for smart cities are increasingly required in our everyday life, looking for efficient management of different services. This manuscript reviews computer vision-based approaches needed for applications, from counting people in crowded scenarios to vehicle classification or free parking slot detection. In recent years, due to the improvements in technology and the increase in computing power, a large number of these approaches have been implemented in real-time or even embedded in low-cost hardware. This chapter starts by proposing a pipeline to perform the review on a common framework. Initially, image processing techniques generally used on urban video analytics are reviewed. Then, object detection and classification approaches are summarized, highlighting that in several recent approaches, both tasks are merged in a single module. Once objects are detected and classified, an important component to be considered is the tracking, which was initially performed by Kalman filters, but recently, deep learning-based approaches have given the best results. Finally, the chapter reviews applications built on top of the modules of the proposed pipeline, as well as datasets generally used as benchmarks for evaluations and comparisons. As mentioned in several sections of the chapter, this work does not pretend to be an exhaustive list of approaches from the state-of-the-art for each module of the proposed pipeline, but rather an illustration of some of the different recent contributions. As a first general conclusion, it could state that deep learning-based solutions are the best option for any of the proposed modules. As a second conclusion, it is clear that there are many opportunities to be explored based on video analytic of urban environments. Depending on the complexity of the solution, low-cost hardware can be considered. Most of the applications reviewed in this chapter are standalone solutions; the challenge for the future is to interconnect these applications to benefit each other and develop more robust solutions for smart cities.

[10] http://www.panda-dataset.com/

Table 1 Papers summarized for each stage of the pipeline and technique used.

Dataset	Preprocessing	Object detection	Object classification	Tracking	Applications
KITTI: [17], [22], [60], [71]	Camera calibration: [16], [23], [68], [72]	Faster R-CNN: [32]	Optical flow: [52]	Kalman Filter: [19]	Pedestrian Detection/Tracking: [1], [29], [56], [58]
PASCAL VOC 2007: [17], [71]	Inverse Perspective Mapping: [33], [49], [55]	Finite State Machine: [27]	Proyective Distortion: [10]	Extended Kalman Filter: [8], [20], [25]	License Plate Recognition: [2], [28], [50], [75]
MS COCO: [9], [71]	Perspective View: [35], [48]	Histogram equalization and RGB-Local Binary Pattern (RGB-LBP): [6]	MLP: [26]	Particle Filter: [5], [29]	Vehicle Counting: [49]
ImageNet: [9], [24]	Background Subtraction: [34], [37], [44], [47]	Optical Flow: [4]	Custom Filtering: [13], [46]	SORT: [35], [64]	Vehicle Classification: [10], [12]
UA-DETRAC: [45]	Mixture of Gaussians (MOG): [62]	CenterNet-3D object detection: [60]	Haar cascade, HOG and SVM: [31], [70]	Custom CNN: [22], [50]	Vehicle Trajectory: [15]
AVSS-2007: [27]	Privacy-preserving strategies: [28], [63]	Custom CNN: [17], [18], [30], [40], [71]	Faster R-CNN: [50]	DeepSORT: [56], [59]	Parking space: [33], [48]
VeRi-776: [50]	Image enhancement: [21], [57], [61], [74]	YOLOv4: [9]	Custom CNN: [24], [43], [65], [66], [73]	3D Tracking: [3], [53], [54]	Face detection: [31], [63]
CDnet: [47]					
Vehicle-1M: [36]					
PANDA: [18]					

Acknowledgements 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

1. Imran Ahmed, Misbah Ahmad, Joel JPC Rodrigues, Gwanggil Jeon, and Sadia Din. A deep learning-based social distance monitoring framework for covid-19. *Sustainable Cities and Society*, 65:102571, 2021.
2. Md Yeasir Arafat, Anis Salwa Mohd Khairuddin, and Raveendran Paramesran. Connected component analysis integrated edge based technique for automatic vehicular license plate recognition framework. *IET Intelligent Transport Systems*, 14(7):712–723, 2020.
3. Ahmad Arinaldi, Jaka Arya Pradana, and Arlan Arventa Gurusinga. Detection and classification of vehicles for traffic video analytics. *Procedia computer science*, 144:259–268, 2018.
4. Sepehr Aslani and Homayoun Mahdavi-Nasab. 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, 2013.
5. Ata-Ur-Rehman, Sameema Tariq, Haroon Farooq, Abdul Jaleel, and Syed Muhammad Wasif. Anomaly detection with particle filtering for online video surveillance. *IEEE Access*, 9:19457–19468, 2021.
6. Danilo Avola, Gian Luca Foresti, Niki Martinel, Christian Micheloni, Daniele Pannone, and Claudio Piciarelli. Aerial video surveillance system for small-scale uav environment monitoring. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.
7. Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.
8. Henk AP Blom and Yaakov Bar-Shalom. The interacting multiple model algorithm for systems with markovian switching coefficients. *IEEE transactions on Automatic Control*, 33(8):780–783, 1988.
9. Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.
10. Biswajit Bose and Eric Grimson. Improving object classification in far-field video. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.
11. Thierry Bouwmans, Sajid Javed, Maryam Sultana, and Soon Ki Jung. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. *Neural Networks*, 117:8–66, 2019.
12. Norbert Buch, Mark Cracknell, James Orwell, and Sergio A Velastin. Vehicle localisation and classification in urban cctv streams. *Proc. 16th ITS WC*, pages 1–8, 2009.
13. Christopher Canel, Thomas Kim, Giulio Zhou, Conglong Li, Hyeontaek Lim, David G Andersen, Michael Kaminsky, and Subramanya R Dulloor. Scaling video analytics on constrained edge nodes. *arXiv preprint arXiv:1905.13536*, 2019.
14. Mingwei Cao, Liping Zheng, Wei Jia, and Xiaoping Liu. Joint 3d reconstruction and object tracking for traffic video analysis under iov environment. *IEEE Transactions on Intelligent Transportation Systems*, 2020.
15. Xianbin Cao, Changxia Wu, Jinhe Lan, Pingkun Yan, and Xuelong Li. Vehicle detection and motion analysis in low-altitude airborne video under urban environment. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(10):1522–1533, 2011.
16. Bruno Caprile and Vincent Torre. Using vanishing points for camera calibration. *International journal of computer vision*, 4(2):127–139, 1990.

17. Ramakant Chandrakar, Rohit Raja, Rohit Miri, Upasana Sinha, Alok Kumar Singh Kushwaha, and Hiral Raja. Enhanced the moving object detection and object tracking for traffic surveillance using rbf-fdlnn and cbf algorithm. *Expert Systems with Applications*, 191:116306, 2022.
18. Kai Chen, Zerun Wang, Xueyang Wang, Dahan Gong, Longlong Yu, Yuchen Guo, and Guiguang Ding. Towards real-time object detection in gigapixel-level video. *Neurocomputing*, 2021.
19. Zezhi Chen, Tim Ellis, and Sergio A Velastin. Vehicle detection, tracking and classification in urban traffic. In *15th Int. Conf. on Intelligent Transportation Systems*, pages 951–956. IEEE, 2012.
20. Hyunggi Cho, Young-Woo Seo, BVK Vijaya Kumar, and Ragunathan Raj Rajkumar. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *Int. Conf. on Robotics and Automation*, pages 1836–1843. IEEE, 2014.
21. Rita Cucchiara, Costantino Grana, Massimo Piccardi, Andrea Prati, and Stefano Sirotti. Improving shadow suppression in moving object detection with hsv color information. In *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 01TH8585)*, pages 334–339. IEEE, 2001.
22. Iván del Pino, Víctor Vaquero, Beatrice Masini, Joan Sola, Francesc Moreno-Noguer, Alberto Sanfeliu, and Juan Andrade-Cetto. Low resolution lidar-based multi-object tracking for driving applications. In *Iberian Robotics conference*, pages 287–298. Springer, 2017.
23. Jonathan Deutscher, Michael Isard, and John MacCormick. Automatic camera calibration from a single manhattan image. In *European Conference on Computer Vision*, pages 175–188. Springer, 2002.
24. Bhaskar Dey and Malay Kumar Kundu. Turning video into traffic data–an application to urban intersection analysis using transfer learning. *IET Image Processing*, 13(4):673–679, 2019.
25. Helgo Dyckmanns, Richard Matthaei, Markus Maurer, Bernd Lichte, Jan Effertz, and Dirk Stüker. 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, 2011.
26. Romuere RV e Silva, Kelson RT Aires, and Rodrigo de MS Veras. Detection of helmets on motorcyclists. *Multimedia Tools and Applications*, 77(5):5659–5683, 2018.
27. Quanfu Fan and Sharath Pankanti. Modeling of temporarily static objects for robust abandoned object detection in urban surveillance. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 36–41. IEEE, 2011.
28. Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in google street view. In *2009 IEEE 12th international conference on computer vision*, pages 2373–2380. IEEE, 2009.
29. Prateek K. Gaddigoudar, Tushar R. Balihalli, Suprith S. Ijantkar, Nalini C. Iyer, and Shruti Maralappanavar. Pedestrian detection and tracking using particle filtering. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 110–115, 2017.
30. Chenqiang Gao, Pei Li, Yajun Zhang, Jiang Liu, and Lan Wang. People counting based on head detection combining adaboost and cnn in crowded surveillance environment. *Neurocomputing*, 208:108–116, 2016. SI: BridgingSemantic.
31. KS Gautam and Senthil Kumar Thangavel. Video analytics-based intelligent surveillance system for smart buildings. *Soft Computing*, 23(8):2813–2837, 2019.
32. Raducu Gavrilescu, Cristian Zet, Cristian Foșalău, Marcin Skoczylas, and David Cotovanu. Faster r-cnn:an approach to real-time object detection. In *2018 International Conference and Exposition on Electrical And Power Engineering (EPE)*, pages 0165–0168, 2018.
33. Giulio Grassi, Kyle Jamieson, Paramvir Bahl, and Giovanni Pau. Parkmaster: An in-vehicle, edge-based video analytics service for detecting open parking spaces in urban environments. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, pages 1–14, 2017.
34. Piotr Graszka. Median mixture model for background–foreground segmentation in video sequences. 2014.

35. Alexander Grents, Vitalii Varkentin, and Nikolay Goryaev. Determining vehicle speed based on video using convolutional neural network. *Transportation Research Procedia*, 50:192–200, 2020.
36. Haiyun Guo, Chaoyang Zhao, Zhiwei Liu, Jinqiao Wang, and Hanqing Lu. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
37. Surendra Gupte, Osama Masoud, Robert FK Martin, and Nikolaos P Papanikolopoulos. Detection and classification of vehicles. *IEEE Transactions on intelligent transportation systems*, 3(1):37–47, 2002.
38. Amal Ben Hamida, Mohamed Koubaa, Chokri Ben Amar, and Henri Nicolas. Toward scalable application-oriented video surveillance systems. In *2014 Science and Information Conference*, pages 384–388. IEEE, 2014.
39. Jean-Philippe Jodoin, Guillaume-Alexandre Bilodeau, and Nicolas Saunier. Urban tracker: Multiple object tracking in urban mixed traffic. In *IEEE Winter Conference on Applications of Computer Vision*, pages 885–892. IEEE, 2014.
40. Mohamad Haniff Junos, Anis Salwa Mohd Khairuddin, and Mahidzal Dahari. Automated object detection on aerial images for limited capacity embedded device using a lightweight cnn model. *Alexandria Engineering Journal*, 2021.
41. R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.
42. Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
43. T Senthil Kumar. Video based traffic forecasting using convolution neural network model and transfer learning techniques. *Journal of Innovative Image Processing (JIIP)*, 2(03):128–134, 2020.
44. B Lee and M Hedley. Background estimation for video surveillance. 2002.
45. Chenge Li, Gregory Dobler, Xin Feng, and Yao Wang. Tracknet: Simultaneous object detection and tracking and its application in traffic video analysis. *arXiv preprint arXiv:1902.01466*, 2019.
46. Yuanqi Li, Arthi Padmanabhan, Pengzhan Zhao, Yufei Wang, Guoqing Harry Xu, and Ravi Netravali. 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, 2020.
47. Kyungsun Lim, Won-Dong Jang, and Chang-Su Kim. Background subtraction using encoder-decoder structured convolutional neural network. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017.
48. Xiao Ling, Jie Sheng, Orlando Baiocchi, Xing Liu, and Matthew E Tolentino. Identifying parking spaces & detecting occupancy using vision-based iot devices. In *2017 Global Internet of Things Summit (GIoTS)*, pages 1–6. IEEE, 2017.
49. Chenghuan Liu, Du Q Huynh, Yuchao Sun, Mark Reynolds, and Steve Atkinson. A vision-based pipeline for vehicle counting, speed estimation, and classification. *IEEE Transactions on Intelligent Transportation Systems*, 2020.
50. Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European conference on computer vision*, pages 869–884. Springer, 2016.
51. Xinyue Liu, Jun Sang, Weiqun Wu, Kai Liu, Qi Liu, and Xiaofeng Xia. Density-aware and background-aware network for crowd counting via multi-task learning. *Pattern Recognition Letters*, 150:221–227, 2021.
52. Alisa Makhmutova, Igor V Anikin, and Maria Dagaeva. Object tracking method for videomonitoring in intelligent transport systems. In *2020 International Russian Automation Conference (RusAutoCon)*, pages 535–540. IEEE, 2020.
53. Umesh Parameshwar Naik, Varadi Rajesh, Rohith Kumar, et al. Implementation of yolov4 algorithm for multiple object detection in image and video dataset using deep learning and

artificial intelligence for urban traffic video surveillance application. In *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–6. IEEE, 2021.
54. Thien-Nghia Nguyen, Bernd Michaelis, Ayoub Al-Hamadi, Michael Tornow, and Marc-Michael Meinecke. Stereo-camera-based urban environment perception using occupancy grid and object tracking. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):154–165, 2011.
55. Byeongjoon Noh, Wonjun No, Jaehong Lee, and David Lee. Vision-based potential pedestrian risk analysis on unsignalized crosswalk using data mining techniques. *Applied Sciences*, 10(3):1057, 2020.
56. SM Praveenkumar, Prakashgouda Patil, and PS Hiremath. 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, 2022.
57. Hongquan Qu, Tongyang Yuan, Zhiyong Sheng, and Yuan Zhang. A pedestrian detection method based on yolov3 model and image enhanced by retinex. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE, 2018.
58. Daniela Ridel, Eike Rehder, Martin Lauer, Christoph Stiller, and Denis Wolf. A literature review on the prediction of pedestrian behavior in urban scenarios. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3105–3112. IEEE, 2018.
59. Adson M Santos, Carmelo JA Bastos-Filho, Alexandre MA Maciel, and Estanislau Lima. Counting vehicle with high-precision in brazilian roads using yolov3 and deep sort. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 69–76. IEEE, 2020.
60. Yuguang Shi, Yu Guo, Zhenqiang Mi, and Xinjie Li. Stereo centernet-based 3d object detection for autonomous driving. *Neurocomputing*, 471:219–229, 2022.
61. Zhenghao Shi, Bin Guo, Minghua Zhao, Changqing Zhang, et al. Nighttime low illumination image enhancement with single image using bright/dark channel prior. *EURASIP Journal on Image and Video Processing*, 2018(1):1–15, 2018.
62. Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE, 1999.
63. Nguyen Anh Tu, Kok-Seng Wong, M Fatih Demirci, Young-Koo Lee, et al. Toward efficient and intelligent video analytics with visual privacy protection for large-scale surveillance. *The Journal of Supercomputing*, pages 1–31, 2021.
64. Henry O. Velesaca, Steven Araujo, Patricia L. Suárez, Ángel Sánchez, and Angel D. Sappa. Off-the-shelf based system for urban environment video analytics. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 459–464, 2020.
65. C Vishnu, Dinesh Singh, C Krishna Mohan, and Sobhan Babu. Detection of motorcyclists without helmet in videos using convolutional neural network. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3036–3041. IEEE, 2017.
66. Cheng Wang, Ming Cheng, Ferdous Sohel, Mohammed Bennamoun, and Jonathan Li. Normalnet: A voxel-based cnn for 3d object classification and retrieval. *Neurocomputing*, 323:139–147, 2019.
67. Haoran Wei, Matthew Laszewski, and Nasser Kehtarnavaz. Deep learning-based person detection and classification for far field video surveillance. In *2018 IEEE 13th Dallas Circuits and Systems Conference (DCAS)*, pages 1–4. IEEE, 2018.
68. Horst Wildenauer and Branislav Micusik. Closed form solution for radial distortion estimation from a single vanishing point. In *BMVC*, volume 1, page 2, 2013.
69. Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.
70. Ronghua Xu, Seyed Yahya Nikouei, Yu Chen, Aleksey Polunchenko, Sejun Song, Chengbin Deng, and Timothy R Faughnan. Real-time human objects tracking for smart surveillance at

the edge. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2018.
71. Hui Zhang, Kunfeng Wang, Yonglin Tian, Chao Gou, and Fei-Yue Wang. Mfr-cnn: Incorporating multi-scale features and global information for traffic object detection. *IEEE Transactions on Vehicular Technology*, 67(9):8019–8030, 2018.
72. Mi Zhang, Jian Yao, Menghan Xia, Kai Li, Yi Zhang, and Yaping Liu. Line-based multi-label energy optimization for fisheye image rectification and calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4137–4145, 2015.
73. Jiasong Zhu, Ke Sun, Sen Jia, Qingquan Li, Xianxu Hou, Weidong Lin, Bozhi Liu, and Guoping Qiu. Urban traffic density estimation based on ultrahigh-resolution uav video and deep neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12):4968–4981, 2018.
74. Alexander Zotin. Fast algorithm of image enhancement based on multi-scale retinex. *Procedia Computer Science*, 131:6–14, 2018.
75. Yongjie Zou, Yongjun Zhang, Jun Yan, Xiaoxu Jiang, Tengjie Huang, Haisheng Fan, and Zhongwei Cui. License plate detection and recognition based on yolov3 and ilprnet. *Signal, Image and Video Processing*, pages 1–8, 2021.