Espinosa, J. E., Velastín, S. A. and Branch, J. W. (2020). Detection of Motorcycles in Urban Traffic Using Video Analysis: A Review. IEEE Transactions on Intelligent Transportation Systems.
DOI: 10.1109/TITS.2020.2997084.

Detection of Motorcycles in Urban Traffic Using Video Analysis: A Review

Jorge E. Espinosa, Sergio A. Velastin, *SMIEEE,* and John W. Branch

***Abstract*—Motorcycles are Vulnerable Road Users (VRU) and as such, in addition to bicycles and pedestrians, they are the traffic actors most affected by accidents in urban areas. Automatic video processing for urban surveillance cameras has the potential to effectively detect and track these road users. The present review focuses on algorithms used for detection and tracking of motorcycles, using the surveillance infrastructure provided by CCTV cameras. Given the importance of results achieved by Deep Learning theory in the field of computer vision, the use of such techniques for detection and tracking of motorcycles is also reviewed. The paper ends by describing the performance measures generally used, publicly available datasets (introducing the Urban Motorbike Dataset (UMD) with quantitative evaluation results for different detectors), discussing the challenges ahead and presenting a set of conclusions with proposed future work in this evolving area.**

***Index Terms*—VRU (Vulnerable Road Users), Motorcycle detection, Vehicle Detection, Tracking, Convolutional Neural Networks (CNNs), deep learning, Computer Vision.**

I. Introduction

THE concentration of the world's population in cities is increasing year by year. It is calculated that 53% of the population currently lives in urban areas and by 2050 this number will grow to 67%. Nowadays, 64% of people travel within urban environments and is expected that by 2050 the amount of kilometers traveled will increase by 300% [1].

Emerging countries face the additional challenge of highly populated urban areas with inadequate and insufficient urban infrastructure [2]. For this reason, their mobility strategies include efforts to stimulate eco-friendlier modes such as public transport, walking, cycling, etc. However, people's limited financial resources and the relative ease to obtain driving licenses, e.g. for motorcycles smaller than 200 c.c., have resulted in a significant increase in this kind of vehicle (Figure 1). Furthermore, there have been adverse consequences in public health, particularly in Latin-America, where fatal casualties involving motorcycles account for 45% of (all) traffic accidents [3] and a significant impact on the environment due to the particle emission of this type of vehicle (P.M. 2.5). There are reports [4] for other emerging regions in the world, such as the Middle East, where they indicate that 63% of traffic accidents involve VRUs, of which 32% corresponds just to motorcyclists .

Fig. 1. Motorcycles as part of traffic jams. Note the overflowing into the dedicated BRT lane and the overall congestion. Photograph taken from [5]

According to the World Health Organization (WHO), in 2018 the proportion of VRUs involved in fatalities in traffic accidents represented more than 54% [6] of all road users. "*Vulnerable road user*" is a term coined to refer to traffic actors more exposed to risk. Pedestrians, cyclists and motorcyclists are accordingly considered "*vulnerable since they benefit from little or no external protective devices that would absorb energy in a collision*" [7].

Therefore, it is important to investigate and evaluate techniques or strategies that allow detecting and tracking pedestrians, bicyclists and motorcycles to reduce accidents, optimizing and improving urban traffic management. "*Of all the entities in the class of vulnerable road users (VRUs), pedestrians, bicyclists and motorcyclists are the most likely to suffer severe injuries and death if they are involved in a collision with an automobile.*"[8]. The use of Intelligent Transportation Systems (ITS), and video analysis in particular, could be one way of dealing with the issues affecting motorcycles in the urban context, specially in emerging countries, where motorcycles constitute an important means of transport.

This review is specially oriented to the detection and tracking of motorcycles since vision-based pedestrian detection has been extensively investigated during recent years, e.g. [9] for ADAS (Advanced Driver Assistance Systems) and [10], [11]. There is also a complete benchmark for vision-based cyclist detection developed by Li et al. in [12]. On the other hand, very limited work has been undertaken on motorcycles, even though approaches might share similar techniques as for

J.E. Espinosa is with Politécnico Colombiano Jaime Isaza Cadavid, Carrera 48 No. 7-151 El Poblado, Medellín - Colombia (e-mail: jeespinosa@elpoli.edu.co).

S.A. Velastin is with Zebra Technologies Corp., UK, Queen Mary University of London, UK and University Carlos III Madrid, Spain (e-mail: sergio.velastin@ieee.org).

J.W. Branch is with Universidad Nacional de Colombia – Sede Medellín, Calle 59 A N 63-20 – Medellín - Colombia (e-mail: jwbranch@unal.edu.co).

pedestrians and cyclists.

Because most researchers working on VRUs tend to follow a somewhat well-established pipeline of processes common in object detection, classification a nd t racking, t his r eview will consider relevant work in the main sub-areas of such pipeline (Figure 2).:

- Detection: referred here to as Hypothesis Generation (HG), encompassing three main aspects:
  - Feature extraction
  - Segmentation
  - Localization.
- Classification or hypothesis verification (HV).
- Tracking.

It is also important to evaluate the performance of the studied algorithms especially under occlusion conditions frequent in congested urban traffic conditions and where detection, classification and tracking are more critical for traffic control centers. Thus, when comparing algorithms, where possible we comment on whether they account for clutter.

Finally, in recent years there have been new approaches based on what are called deep learning methods that therefore deserve a special attention. The most applied technique for detection is known as Convolutional Neural Networks (CNN), so detection of motorcycles using CNNs is especially considered. The final section considers the main challenges that motorcyclist detection faces, listing some of the public datasets available for this type of research and the performance measures used to evaluate algorithms. A new dataset focusing on motorcycles is introduced, presenting baseline results. The paper ends by pointing out some topics that are still open for further research in this field.

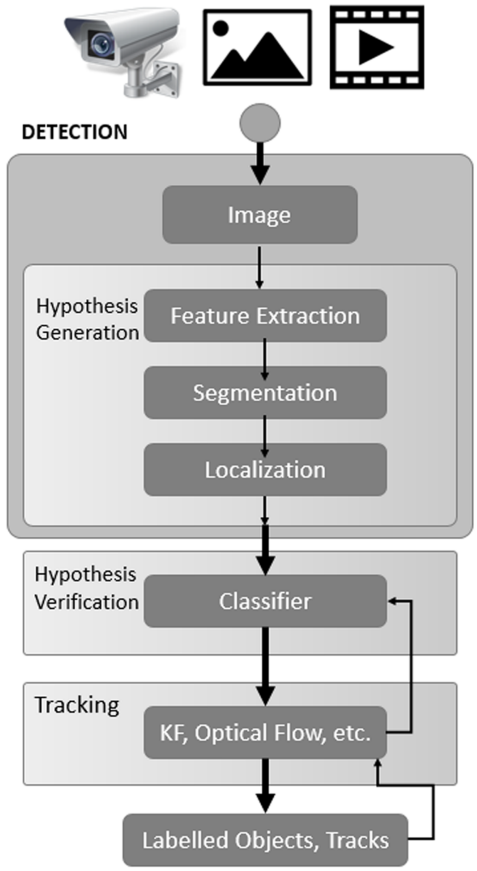

Fig. 2. Traditional sequence of activities used for VRU detection and tracking

II. Hypothesis Generation

Hypothesis Generation (HG) for motorcycle detections approaches can be divided into two categories: appearance-based (e.g. using shape, color and texture methods) and motion-based that use dynamic, spatio-temporal characteristics methods. While appearance-based methods tend identify the object directly from single images, motion-based methods requires a sequence of images (video) to extract the dynamic and static characteristics and to determine the possible objects present in the scene. In the context of vehicle detection, both methods try to separate road users from the background. As described later, deep learning methods can overcome the dependency on manually choosing what particular image features to use.

A. Methods based on appearance:

Appearance-based approaches detect motorcycles directly from the image, detecting parts of the object to be classified and used to identify an object. Appearance features are obtained from visual information extracted from objects including color, texture or shape. Using these features it is possible to detect vehicles even if remain static as in a traffic jam.

1) Explicit shape approaches: As a first approximation based on shape features, the spatial relationships between parts of objects detected is modelled by explicit shape approaches. For instance, the Circular Hough Transform (CHT) is used so that the edge points of the potential objects (tires, helmets) are grouped into object candidates by a voting procedure over a set of parameterized image objects. This technique is used by Silva et al. [13], [14] for helmet detection in motorcycles, useful for ROI (Region of Interest) localization. This strategy is also used for helmet and headlight detection by Mukhtar and Tang [15]. Nevertheless the method is prone to errors since it generates many false positives in congested scenarios and detection fails for affine or projective transformations.

2) Texture features: Feature extraction and description advances go from basic image features (e.g. edge and symmetry) to more general and most robust features that are invariant to illumination, object size, rotation and affine transformations changes. Scale-invariant feature transform (SIFT) is a feature descriptor that produces local features which are invariant to image scaling, rotation and translation and that are partially invariant to affine projections and illumination changes [16]. These features identify the presence of salient points, which are robust to many geometrical transformations and even to illumination changes. In some approaches, SIFT is modified as in DSIFT (Dense Scale Invariant Feature Transform) [17], whose main property is that the features are generated in a uniform way based on a dense mesh of places where scale and orientation are fixed. The speeded up robust features (SURF), is also inspired on the SIFT descriptor but it focuses

more on computational speed, using box filters instead of Gaussian filters. Its feature descriptor is constructed with the sum of the Haar wavelet response around the interest point. However, none of these three feature extractors incorporate spatial relationships between key points. For this reason Thai et al. [18] adopt a technique of pyramidal space kernel for feature extraction [19], exploiting the spatial localization of motorcycles in their training images. The method relies heavily on images captured from a top view angle which constrains the field of view of a typical surveillance camera.

The Histogram of Oriented Gradients (HOG) extracts features based on a combination of color and texture, which are extracted by first applying edge operators over the image and then discretizing and bucketing the orientations of the edge intensities into a histogram achieving some spatial and illumination invariance. Due to the high vulnerability of motorcyclists, many motorcycle detection approaches also analyze the presence of helmets. The use of HOG is quite common for this. A feature comparative analysis is reported by Dahiya et al. [20] and Singh et al. [21], comparing the performances of three different appearance features obtained after foreground detection: HOG, SIFT and LBP. HOG shows the best discrimination properties when used both for motorcycling and helmet detection. The same strategy is proposed by [22], but without reporting results. Since such methods rely on foreground object detection, it is no clear how they could deal with occluded or stationary objects. The HOG descriptor combined with the Circular Hough Transform (CHT) is also used for helmet detection in [14], since geometric information obtained only from CHT may be ambiguous w.r.t heads shapes. Recently, exploiting the arc circularity detected in ROI areas with a combination with HOG features, Talaulikar et al. [23] detect helmets on motorcycle riders, applying Principal Component Analysis (PCA) for performance and accuracy improvements, but they provide very little details on the statistical significance of their results and the applicability to surveillance cameras given their reliance on background removal. Meanwhile Baris and Bastanlar [24] combine an omnidirectional and a Pan-Tilt-Zoom (PTZ) camera in a hybrid camera system to detect and track traffic scene objects including pedestrians, motorcycles, cars and vans. Initial classification is done using an omni-directional camera using shape-based features of moving objects extracted by Adaptive Background Learning (ABL) (see section II-B). This classification is improved once the moving objects are detected and tracked using a Kalman filter. Then, from PTZ camera images the object HOG features are extracted and used to increase classification accuracy. The method is very sensitive to occlusions or overlapping objects, where the background subtraction method may fail.

Computing the difference of the sum of pixels within rectangles in a given image patch, Haar-like features allow real-time performance and are used for motorcycle detection to improve vehicle driving safety in [25] using it for helmet detection in [26], [27]. Nevertheless, Haar-like features do not correlate under different view angles, restricting their use in surveillance scenarios.

3) Geometric features & 3-D Models : Computer-generated 3-D models of motorcycles can be used for detection by appearance matching. Messelodi et al. [28] extend detection and classification of vehicles in urban intersections [29], achieving discrimination between bicycles and motorcycles. An object descriptor stores features to decide which view is selected to analyze the image. Two different contexts are identified based on the movement direction of the vehicle with respect to the axis orthogonal to the ground plane. Small angles describe front or rear view of the vehicle appearing in the image. Otherwise, the image represents a lateral view of the vehicle (Figure 3). In a front view, the wheel section will determine a comparative criterion between the whole of the object with respect to the thickness of the rim, distinguishing between a bicycle and a motorcycle. Unfortunately, the algorithm is not evaluated under congested scenarios were occlusions could affect the feature extraction process. Meanwhile, Buch et al. [30] use a 3-D wire model to discriminate urban road users between bus/lorry, van, car/taxi, motorcycle/bicycle and pedestrian. The approach does not take into account the movement of the vehicles and does not deal with occlusions as each computed silhouette is associated with a single vehicle.

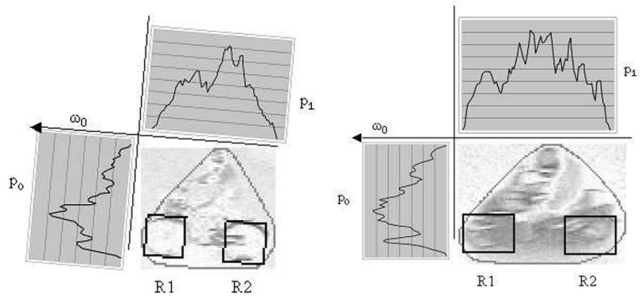

Fig. 3. Side View: The luminance of the region inside the wheels of a bicycle is more similar to the background, with respect to the same region of a motorcycle. Based on the parameters extracted, the two wheels are used as criteria for classification (reprinted from [28], by permission from Elsevier).

4) Multiple features: Due to the complexity of motorcycle detection, some authors combine multiple features to solve problems such as occlusion, background clutter and variations in orientation and illumination. Silva et al. [13], [14] proposed a methodology for helmet use detection on highways scenarios by determining which objects appear to be moving using a calibration stage, extracting their features by SURF, HAAR, HOG, Fourier and K-means description. Even with the improved results reported by the authors, the use of multiple integrated features demands a complex task of parameter selection and tuning and applicability for urban scenarios is not clear.

5) Other descriptors: There are other approaches to extract motorcycle features described in the literature. Le and Huynh [31] propose an integrated method for counting and detecting motorcycles, extracting features with Gabor filters and using random forest to generate a density map via an indirect counting method. Using interest point descriptors, Muzammel et al. [32] do hypothesis generation for a motorcycle vision-based rear-end collision detection system using Harris corners which are points uniformly sampled in a salience detector. However, this procedure could be sensitive to what is known as the "aperture problem". Illumination variances can be removed

by edge detection and binary edges can provide normalized input for feature descriptors. So, the Sobel operator has been used [33] to design collision avoidance systems, detecting the rear view of motorcycles, by defining i nitial hypotheses of possible objects from shadows, wheels and vertical edges [34] and using template matching [35]. However, these approaches are not suitable for CCTV surveillance cameras due to a restricted distance and view angle (motorcycle rear view). Duan et al. [34] monitored the back of vehicles and motorcycles supported by a Lane Change Assistant (LCA). A multi-resolution technique is presented to achieve real-time performance, using an Integrated Memory Array Processor (IMAP). For HG, the method involves obstacle detection using optical flow a nd p rior k nowledge u sing d ata f rom the detection of previous vehicles during the day that includes symmetry, color, shadows, geometric characteristics (corners, horizontal and vertical edges) and texture. It is not clear how prior data can be sufficiently r epresentative f or a ll possible detections. Initial hypotheses of possible objects are obtained from shadows, wheels and vertical edges. As shadows are usually affected by illumination, the ROI position is adjusted based on grayscale symmetry. The main issue with the method is related to dealing with congested scenes with three or more objects overlapping, or reduced light conditions.

B. Methods based on Motion:

Using spatio-temporal information for detection is popular, because vehicle motion can be used for traffic counts, infringements, etc. Motion detection aims to separate moving foreground objects from the static background. Generally when the background is modeled with information of the scene, then moving objects are detected by comparing the current frame with the background model. Due to dynamics, speed and maneuverability of motorcycles, most of the works rely on methods based on motion features using background subtraction, nevertheless further analysis is required to discriminate precisely this vehicle type. Furthermore, working under congestion and semi-static traffic makes it difficult to obtain good background models.

1) Simple Background subtraction: The Background model is initially defined (or learned) with no movement objects (traffic) as a reference scene. Each frame is subtracted from the background model applying a threshold for the resulting difference, generating a foreground mask. Such threshold can be dynamic or constant, global or local. This approach is used by Kanhere et al. [36] for motorcycle detection, starting with a manual "six clicks" camera calibration, defining three lines parallel to the image projection, then applying background subtraction to identify blobs of moving objects. Finally, features are identified using the estimated height and the local slope (in real world coordinates) projected from foreground blobs and grouped by region growth. These features correspond to typical vehicle dimensions that depend on camera calibration. It is not clear if these features can be identified robustly under heavy traffic congestion. Meanwhile, the work of Chiu et al. [37] is extended by Ku et al. [38], where hypothesis generation of motorcycle detections starts from background subtraction, dealing with possible overlaps with other vehicles and proposing an additional segmentation method to detect and separate motorcycles in the scene. The methods rely on helmet detection to drive search methods. Nevertheless, when people do not wear helmets it makes them more vulnerable, so it is even more important to detect them. To deal with this situation, Chiverton [39] shows a method for detecting motorcycle riders who do not wear helmets. The method is based on background subtraction and component labeling, along with operations to reduce noise and to add pixels to detected regions. Once a motorcycle is detected, a search process is employed to find the rider's head. However, it is not clear if they can deal with overlapping objects (traffic congestion) and not detected as a single blob. Background subtraction is also used by Sekar et al. [40] for motorcycle features extraction later used for helmet detection.

2) Gaussian Mixture Models (GMM): [41] is the most used technique for foreground extraction. It consists on temporally modelling each pixel of the image as a mixture of two or more Gaussians, updating the distributions according to the temporal pixel changes. This has a high computational cost, but it can deal up to some extent with illumination changes and frequent clutter. Waranusast [42] detects moving objects using GMM models and morphological closing on the resulting binary images. To reduce computational burden, only one instance of the frame sequence is captured in front of the camera, this also allows determination of the direction of the moving object. This technique works for fluid traffic conditions, with no overlapping of vehicles where blob detection does not merge objects. Meanwhile, Dupuis et al. [43] subtract background using a combination of GMM and a shadow removal strategy. Then, the foreground blobs are refined through Laplacian densities, dealing with different types of shadows in the scene. Thereafter, simple features such as area, width and height are employed for discriminating motorcycles from other vehicles. Rashidan et al. [44] classify common moving objects in street scenes as motorcycles, pedestrians and cars. Moving objects are detected using GMM, extracting foreground object features in terms of spatial and temporal attributes. Spatial properties are based on a criterion of "compactness" (ratio between area and square perimeter) and height to width ratio. Temporal attributes are obtained also from optical flow, working only with the central point of the detected objects to reduce computational burden.

Traditional GMM has been improved to deal with shadow detection and lighting changes. Chen et al. [45] propose a vehicle (including motorcycles) detection and tracking system from roadside CCTV in urban traffic. The Multi-Dimensional Gaussian Kernel density Transform (MDGKT) has been used in an attempt to deal with unwanted motions as well as an extension of GMM called Self Adaptive GMM [46] to continuously update the parameters of a GMM. It can also detect object shadows based on the chromaticity value of a moving pixel detected as a foreground. Nevertheless, the method still has problems in detecting overlapping objects moving at similar speeds. Meanwhile, Dahiya et al. [20] and Singh et al. [21] report an approach for HG that relies on adaptive background modelling based on GMM to separate

moving objects. The authors claim that a modified version deals with different lighting conditions, but it is not clear how to determine dynamic parameters such as the learning rate, the lighting threshold or the number of Gaussians. In the hybrid camera system of [24], the Adaptive Background Learning (ABL) is used by an omni-directional camera and an improved GMM (MOG) in a PTZ camera, to detect moving objects. However, methods based on motion generally assume that the videos are captured from a stationary camera and therefore have issues with slow-moving/stopped traffic (traffic jams), abrupt changes in illuminations, noise produced by windy conditions or urban furniture as waving trees and, of course, moving cameras.

III. Hypothesis Verification

Hypothesis verification (HV) is the process of classifying the detected objects (into pre-determined classes such as "pedestrian", "bicycles", "motorcycle", etc.), normally using a supervised learning approach. Obviously, the use of supervised learning generally involves procurement and labelling of possibly large amounts of data.

Sometimes the problem is considered as a binary classification exercise between the object to identify and the background. In other cases, the classes are defined explicitly and the discrimination task is to determine which class a detected object belongs to. Classifiers can be divided in two categories, discriminative and generative classifiers. Discriminative classifiers are the most reported in literature and they learn decision boundaries between classes (e.g. object vs background). Generative classifiers, learn the underlying class distribution detected and are less common for vehicle detection. In what follows, different techniques are reported applied to motorcycles.

A. Discriminative classifiers

These include artificial neural networks (ANNs), support vector machines (SVMs), boosting and conditional random fields (CRFs) commonly treating motorcycle classification as a binary classification problem, vehicle or non-vehicle, because authors tend to focus on one of those vehicles.

1) Support Vector Machines (SVM): SVM carries out classification using linear decision hyperplanes in the feature space. During training, the hyperplanes are calculated to maximize the separation of the training data with different labels [47]. Vectors where the hyperplane lie are called support vectors. The process can be analyzed as a quadratic optimization, producing an absolute optimum. This technique has been used for motorcycle classification, where SVMs (with linear an non-linear kernels) have proved popular. Mukhtar and Tang [15] use HOG features of motorcycles and background to train a linear SVM. No overlapping motorcycles were analyzed. Sekar et al. [40] identify helmet presence by using SVM as a classifier of the visual features extracted from the object. Meanwhile, Duan et al. [34] report a binary SVM classification by hierarchies, first between vehicles and not vehicles, then on vehicle category, it tries to differentiate between motorcycles and not motorcycles. The method has issues dealing with occlusion and under special weather conditions as rain, clouds and night time scenes. Using histograms obtained from head regions (identified by their brightness characteristics due to helmets) Chiverton [39] train a linear SVM. It is not clear how the method deals with the detection of two occupants in a single motorcycle, nor is it tested in congested scenarios. Chen et al. [45] do vehicle detection by training an SVM classifier with synthetic data, with a feature vector of 202 elements comprising measurement based features (MBF) and intensity pyramid-based HOG (IPHOG). The method classifies different types of vehicles, including motorcycles, under different illumination and weather conditions. Nevertheless the feature definition requires high computational resources, which are not evaluated on parallel or GPU architectures. An on-board camera is studied by Shuo and Choi [25], who use SVM as a classifier, but which fails to detect motorcycles due to their ambiguous appearance w.r.t. pedestrians, using Haar-like features as a feature extraction method. An SVM kernel comparison is done by Dahiya et al. [20] and Singh et al. [21] using sigmoid and radial basis function (RBF) and evaluating three types of features (HOG, SIFT and LBP) to arrive at the best hyper-plane results. This binary classification is performed for motorcycles and for helmet presence. Best results were obtained using HOG as a feature with a linear kernel. The main drawbacks are reliance on background subtraction and lack of robustness against occlusions in congested traffic scenes.

SVMs are also used in conjunction with 3D-Models: Messelodi et al. [28] use models of vehicles compared against a set of 3-D definitions. They use 3-D classifiers which calculated best correspondence with each 3-D model performing direct comparisons. Otherwise, an SVM classifier models the difference between motorcycles and bicycles. Results are reported using their own dataset, not publicly available.

SVMs are used used in the hybrid camera system of [24] using HOG features and two binary SVMs, one to discriminate car/van classes and another for motorcycles/pedestrian classes. Best results are obtained from the previous results of the omnidireccional camera classification, which are refined by the SVM classifiers. Final results obtain an accuracy of 98.59%, with only 71 motorcycles to classify.

The potential drawbacks of using an SVM as a classifier are related to the "curse of dimensionality" when there is a kernel projection to a higher order making the learning process slow.

2) Decision trees: Decision trees are useful to model the decision path which leads to vehicle classification. For motorcycle classification, in [43] all detected blobs are manually labeled and fed to a decision tree (pruned to reduce overfitting) obtaining the classification rules. However, these rules are specific to a given camera resolution, lens focal length and camera position.

3) Random Forest: is an ensemble learning method which combines multiple decision trees during training time and outputting the most frequently classified class of the individual trees. They are intended to correct the overfitting of individual decision trees. Le and Huynh [31] use Random Forest to generate a map of features, then an SVM is used to classify the different parts of the motorcycle. They report

high accuracy mainly due to the use of a top-down camera view point, avoiding occlusion between vehicles. Meanwhile [48] uses a patch-based random forest of local experts, used to generate a different object configuration later assembled for final classification.

4) K-nearest neighbors (k-NN): Is a classifier that assigns a label class to the input closest to the k training examples in the feature space. The membership class assignment is based on a simple voting mechanism. When k = 1, the class assignment corresponds to the single nearest neighbor. k-NN is used for motorcycles discrimination and helmet detection in [42]. The method creates features from the geometrical relationship of motorcycles and helmets. The approach relies highly on the quality of the background subtraction for motorcycle and head detection which, as mentioned earlier, and it would have difficulties under clutter or for semi-static or stopped vehicles. k-NN is also used in the hybrid camera system of [24] for the omni-directional camera, using k=5 and obtaining a 100% of accuracy on a relatively small dataset with only 71 motorcycles.

5) Artificial Neural Networks (ANN's): Inspired on the synapses connection of neurons in the human brain, the model consists on artificial neurons and signal connections. Each connection transmits a signal from neuron to neuron which inhibits or activates a possible response. The data transmitted through the network allows the ANN structure to learn the relationship or mapping between inputs and outputs, allowing it to find patterns in the input data. Their use has been reported for motorcycle classification even though their training features requires many parameters to tune and may not converge to a local optimum. On the other hand, results on Deep Learning architectures are raising the interest of the research community (see section V). Sutikno et al. [49] report a classification process of motorcyclists wearing or not helmets, from images captured on the highway using a backpropagation neural network. It is not clear how the network architecture is defined. The training process uses only 150 images previously segmented and it is unclear if this is enough for generalization. It then uses just 30 images restricted to the head area. No details are provided about the stopping criteria for learning, nor on their object detection process. Silva et al. [13], [14] report helmet detection, feeding a Multi-Layer Perceptron (MLP) using ROI definition and circular Hough transform in combination with HOG. Images where obtained from a highway scenario with no congestion nor occlusion and under good weather and light conditions.

B. Other approaches

Fuzzy logic approaches are also used in motorcycle classification. Rashidan et al. [44] use an Adaptive Neuro-Fuzzy Inference System (ANFIS) classifier, based on the first-order Takagi-Sugeno-Kang (TSK) method, achieving improved comparative results for motorcycles and cars. However, the sequences used for comparative proposes contain at most only 1034 frames, having only three motorcycles for detection.

Some work does not use classifiers e.g. [32] and [35] where a bounding box is used for classification, which is constructed based on width localization of the edge that corresponds to the lower part and upper part of the vehicle. The bounding box size and ratio corresponds to a constant parameter defined a priori. Although the authors report work with different resolutions, it is not clear why poor results are obtained with higher resolution images. No clear performance results are presented nor a description of the dataset used.

Matching techniques are used in conjunction with tracking for motorcycle classification in cluttered scenarios in [36], where classification starts once an object is tracked for five consecutive frames. The foreground mask is compared by means of a transformation matrix against real world parameters. To identify the vehicles, a correspondence and coincidence (matching) process is applied. The data used corresponds to more than 2000 vehicles, filmed during a rally in two locations at Mertye Beach (South Carolina). It is not clear how the measurement strategy based on camera calibration allows distinguishing more than one or three vehicles detected as a single blob by background subtraction.

IV. Tracking

In its simplest definition, tracking can be understood as estimating the trajectory of an object in the image plane as it moves around a scene [54], [50] and it involves locating the target in subsequent video frames after it has been recognized and classified by the HV step. It is used to predict vehicle positions in subsequent frames, match vehicles between adjacent frames and ultimately to obtain the trajectory and location of each vehicle for each frame. Some techniques are used to extract vehicles' dynamic attributes, including velocity, direction of movement and vehicle trajectories. Most vehicle tracking algorithms follow one simple principle: vehicles in two adjacent frames are the same if the spatial distance is small.

According to Yilmaz et al. [54], object tracking can be classified as point tracking, kernel based tracking and silhouette based tracking. Point trackers involve detection in every frame, while geometric area or kernel based tracking or contours-based tracking require detection only when the object first appears in the scene. What follows are some of the tracking techniques applied for motorcycles.

A. Kalman Filter Tracking

The Kalman Filter (KF) is a Point Tracker also known as linear quadratic estimation. This algorithm uses a series of measurements observed over time, containing noise and other inaccuracies and produces estimates of unknown variables that tend to be more precise than those based on a single measurement alone. The KF can make full use of the historical information and reduce the search range of the image, to significantly improve system processing speed. Its accuracy and stability can suffer when vehicle motion and light conditions change suddenly, in occlusion conditions or when linearity and assumptions on target dynamics and noise are not satisfied.

Motorcycle tracking in [39] combines a KF for movement estimation and a comparison of photometric data for correspondence analysis. To deal with occlusion and sporadic

TABLE I
MOTORCYCLE DETECTION AND TRACKING ALGORITHMS

Algorithms and Ref.	Features	Cluttering/ Oclussion	Classifier	Tracking	Dataset	Performance
Counting motorcycles [31]	Gabor + Random Forest	High	SVM	N/A	Proprietary data (Top View)	0.9 +/- 0.09 Accuracy
Helmet Detection [20], [21]	HOG, SIFT, LBP	Low	SVM	N/A	Property data	0.98 Accuracy on Motorcycle detection 0.93 Accuracy on Helmet detection
BoVW for motorcycle Detection [18]	SIFT, DSIFT, SURF,P-SIFT, P-SURF	High	BoVW + SVM (RBF kernel)	Blob Tracking for detection	Proprietary data	0.94 F-Score using SURF+ spatial pyramid kernel
HOG for motorcycle [15]	Hough for Circular shapes + HOG	Low	SVM	N/A	Proprietary data	0.96 detection rate
Distance alert system [25]	Haar-like features	Low	SVM	Covariance Tracking	Proprietary data	FP:3/FN:15 Pedestrian FP:0/FN:0 Motorcycles FP:24/FN:28 Cars
Collision Alert [32]	Harris corners	Low	N/A	N/A	Proprietary data – LISA [50] iRoads	TPR 0.954 LISA-dense 0.95 LISA-Urban
Real Time on Road Vehicle Detection [34]	geometric features and texture + wheels contour	Low	SVM	Tracking Window [51]	Proprietary data	0.9173 Average Detection Rate
Edge tracking [35]	Geometrical Features	Low	N/A	Not clearly specified	Proprietary data	Not clearly reported
Helmet Detection [13], [14]	HOG + CHT	Low	Random Forest + MLP	N/A	Proprietary data	0.97 Accuracy for vehicle detection 0.91 for Helmet Detection
SCOCA v2 [28]	3-D Models + Multiple Features	Low	Non-Linear SVM	Kalman predictive filtering technique	Proprietary data	Successful classification rate of 0.967
3-D Models [30]	IM Image measured Features	Low	Measured dimensions (Implies Camera Calibration)	Kalman Filter used for vehicle labelling	i-LIDS datasets	Recall 0.87 Precision 0.85
Chiu et al.[37] and Ku et al.[38]	visual length, visual width, pixel ratio and helmet shape	High	Connected Component Labeling [52]	Velocity and displacement	Proprietary data	Successful detection 0.96 Day 0.80 Night
Helmet presence [39]	HOG Derived from head zone	Low	Lineal SVM	Correspondence analysis and Kalman filters	Proprietary data	Helmet detection Accuracy 0.96 Helmet classification 0.85
Motorcycle Detection During Special Events [36]	Stable Features	Medium	Measured dimensions (Implies Camera Calibration)	Correspondence and matching	Proprietary data	0.04 Error rate Vehicles 0.06 Error rate Motorcycles
Helmet detection [42]	Geometrical Features	Low	K-Nearest Neighbor (KNN)	N/A	Proprietary data	Correct detection rates: 0.84 near lane 0.68 far lane 0.74 both lanes
Overhead Real Time motorcycle Counting [43]	area, height, width	Low	Decision tree	Blob Tracking for detection	Proprietary data	0.058 WEA (Weighted Absolut Error)
NeuroFuzzy detector[44]]	Spatial and Temporalattributes	Low	Fuzzy neural network (FNN)	N/A	Proprietary data + KOGS-IAKS	0.86 Pedestrian 0.88 Motorcycles 0.91 Cars
AutoVDCS [45], [53]	MBF+IPHOG Measurement Based Feature + intensity pyramid-based HOG	Low	SVM	Kalman Filter	Proprietary data + Synthetic Data (Training)	DR 0.96 0.01 FP - 0.05 FN Class. accuracy 0.94.
Helmet detection [49]	Neural weights	Low	MLP	N/A	Property data	0.86 Accuracy Rate
PCA for Helmet detection [23]	HOG + Circular Shape	Low	SVM, MLP, LR	N/A	Property data	0.95 Accuracy LR 0.94 Accuracy MLP 0.95 Accuracy SVM
Helmet detection CCTV [26]	Haar-Features	Low	N/A	N/A	Property data	0.81 Accuracy
Classification & tracking on Hybrid Camera [24]	HOG	Low	SVM	Kalman Filter	Property data	98.59 Accuracy

errors, a "track propagation" strategy is proposed, but it is not clear how this technique could manage multiple parallel tracks or tracks that cross each other, something that can often occur in cluttered scenes. In [45], tracking is done using a Kalman filter, based on centroid location and velocity. For each frame, a class label is computed and the final label for the track is assigned with a voting scheme, considering the entire track to make a decision. As it is only used for class labelling, no specific tracking results are given. A KF is also used in [43], for tracking detected blobs. Results are reported on highways under various traffic and lighting conditions, but these experiments use proprietary and unpublished data and may fail under denser urban traffic conditions.

B. Kanade–Lucas–Tomasi (KLT) feature tracker

KLT is used for two wheelers classification in [55], based on feature extraction from trajectories. Feature-based tracking avoids tracking the moving objects as single bodies, but instead traces distinct features (e.g. Harris corners [56]) of the moving object. Since no descriptive physical features are used for classification and tracking is based on speed, pedaling process and acceleration profiles, the tracker misclassifies e-bikes due

to their higher speeds. Moreover, the method relies heavily on a previous camera calibration and there is no comparison with other classification o r t racking m ethods.The tracking results are not quantified. K LT i s a lso u sed f or Motorcycle detection in [57] using a multi-level motion pattern learning (MLMP) framework for trajectory behavior analysis on video surveillance cameras. That work is based on an open source tool for video traffic a nalysis [58]. H owever, t he n umber of trajectories used to represent a single object could fail under congested scenarios or when using different camera angles (not necessary orthogonal to the objects).

C. Other methods

Other methods for tracking motorcycles are also worth mentioning. A tracking strategy is presented in [28] which combines frame differencing and tracking of edges and corners. The method, however, is not able to deal with overlapping or occluded objects often seen in urban traffic scenarios. In [37], [38] the helmet center of mass is used as a reference point plus velocity and displacement force for predicting the position of a motorcycle in the tracking process. In speed evaluation, results achieved a precision of +/-5 km/hr., comparable with LIDAR or laser speed detection systems. The system relies on a "multi-helmet detection method" when two motorcycles are overlapping. However, it is not clear how it may deal with passengers. Detection by tracking is a method that employs a tracking process to improve detection. In [34] for each object detected, a tracker is initialized and it creates a monitoring window according to the size and position of the object. When a new frame is processed, the tracking mechanism is adapted to the size and position of the tracking window based on its movement estimation [59]. The tracking system can predict the position of an object, according to its speed and historical position, needing the use of a localization technique. Tracking performance is not evaluated with a specific metric, nor evaluated under occluded scenes. Finally, Shuo and Choi [25] perform tracking based on a list of target objects and the criteria for matching is based on histograms, which could fail if there are overlapping detections, specially when the proposed system is used on board of a vehicle.

Table I identifies the main algorithms used for detection and tracking of motorcycles respectively. The third column corresponds to a cluttering (occlusion) factor defined based on the KITTI Vision Benchmark suite [60] as (values in brackets are Maximum Truncation): Low: Fully visible (15%). Medium: Partly occluded (30%). High: Difficult to see (50%). It is important to note the variability in metrics and datasets employed in each algorithm (columns 5 and 6), making it difficult to establish a standardized baseline for algorithm performance comparison.

V. Deep Learning

Deep learning (DL) methods have produced a revolution in the field of computer vision. For pattern recognition the techniques have shown robustness in classifications tasks, being able to deal with different ranges of transformations or distortions as noise, scale, rotation, displacement, illuminance variance, etc. [61]. In object recognition, feature representation obtained from DL often outperforms popular features such as LBP, SURF and HOG [62], [52], [51].

Recent advances on DL achieve object detection in an image through one of two main methods [63]. In the first case through a region proposal based detector, with has as a first stage a region proposal network (RPN) to generate regions of interest and then a second stage where the regions proposed by the RPN are used for bounding box regression and object classification.

The other approach is regression/classification based on Single Stage Detectors (SSDs), which achieve object detection as a regression problem, analyzing the input image and learning the class probabilities and bounding box coordinates. These models can have issues with the detection of small objects or with objects that appear too close in the image, but such single shot architectures can produce real time detection results.

A. Region proposal based detectors

Faster R-CNN [64] is one of the most popular architectures for vehicle detection (but this field moves rapidly). This architecture evolved from R-CNN [65] and Fast R-CNN [66], combining features of a fully convolutional network to perform both region proposals and object detection. R-CNN [65], combines a selective search [67] algorithm for region proposal (RPN) with CNN features to perform object detection. This model was used in [68] to classify motorcycles according to the USA's Federal Highway Administration (FHWA) scheme, reporting 100% precision and 89% recall, but only with 16 motorcycles to detect recorded in a highway without any occlusion. R-CNN generates around 2000 proposal per image and for each proposal a convolutional operation is performed for later classification, making the detection procedure quite slow. Fast R-CNN [66], accelerates the detection process through a single convolution feature map which is generated from the entire image and using a Region of Interest (ROI) pooling layer such as different input sizes can be fed to the classification step.

Methods based on Fast R-CNN approach region proposal use a selective search [67] which is a bottom-up method that iteratively groups small segments of the image based on similarity. This process is computationally expensive and becomes the model bottleneck.

B. Regression/Classification based detectors

Focusing on speed performance (number of frames analyzed per second) at the expense of accuracy, single stage methods achieve detection without proposals. A single convolutional architecture simultaneously predicts bounding boxes and associated class scores. The Single Shot MultiBox Detector (SSD) [69] model consists of a base classification network (in their case the VGG network [70]) for region proposal, suitable for multiple scales, plus a set of convolutional filters used to produce class scores and to determine the bounding box positions. Finally, non-maximal suppression (NMS) [71] is used to eliminate redundant detections. Due to its speed, it is mostly used in autonomous vehicles and advanced driver

assistance systems (ADAS). A single-stage detector reported in [72] uses a CNN for motorcycle detection using only 5 convolutional layers. The last convolutional layer, with a depth of four, splits the negative class (background) into three different classes, claimed to ease learning, given that the negative class will encompass many different features difficult to group into one single class. Redundant detections are eliminated by non-maximal suppression. The model achieves an F1-score of 81% in a video dataset recorded from a top view, which significantly reduces the occlusion of objects.

C. Other approaches

Some methods perform background subtraction for object individualization, then CNN is applied to extract features from the detected moving objects, these features are later used for classification. Such a method is reported for vehicle classification in [73], applying the same strategy for feature extraction in [74], using AlexNet [52] as a feature extractor for motorcycle classification in urban scenarios. Features extracted by the CNN model are then classified by a linear SVM, reaching an almost perfect accuracy on classification (albeit in a small dataset). The method fails when the background subtraction is not robust enough to individualize each vehicle.

A model for detection of motorcyclists without helmet is proposed in [75]. The bounding box for object detection is provided by a GMM model. There is no special consideration for overlapping moving objects that move at similar speeds and could be mixed in the detection. The moving objects are resized to a fixed size to be passed on to the CNN model. No special strategy such as pyramids is used to deal with varying resolution. The CNN model discriminates the detected moving objects as either motorcycles or background, extracting discriminative features from the CNN model and used to perform classification. Finally, the recognition of motorcyclists without helmets is done by cropping the region of the motorcyclist head (which may be highly dependent on the perspective view) and fed to another CNN model which performs binary classification, according to the features trained from motorcyclist heads. Helmet detection is also investigated in [76]. The motorcycle detector uses a linear SVM for a feature vector classification based on histograms of oriented gradienta (HOG). It is not clear how the method identifies the motorcycle nor the region expected to contain the rider's head, which is passed to a CNN for helmet/no-helmet classification. Helmet violations are further processed for location of the license plate using a Haar cascade detector for number plate recognition. A deep learning-based helmet wearing analysis is also proposed in [77] combining GMM for foreground object segmentation and Faster R-CNN for motorcycle detection and helmet presence. The method also includes a license plate recognition to issue fines for traffic violations.

Table II identifies the main deep learning algorithms used for detection of motorcycles.

D. Deep Visual Tracking

Deep learning strategies have been shown to improve the observation model that depicts the appearance of detected objects and are thus potentially useful for tracking purposes. Li et al. [78] provide a review with experimental comparisons of different deep learning trackers, with important conclusions including that the usage of CNN models can significantly improve tracking performance and that deep visual trackers using end-to-end networks usually perform better than trackers that merely using feature extraction networks. The recent work of Chen et al. [79] gives a review of deep learning Multiple Object Tracking (MOT), categorizing, analyzing and comparing deep learning MOT methods. The categories are: 1- Multi-object tracking enhancement using deep network features, 2- Multi-object tracking with deep network embedding and 3- Multi-object tracking with end-to-end deep neural network learning. This categorization is useful to understand the following deep learning methods oriented to motorcycles tracking.

Methods which exploit the use of deep learning features for realizing MOT (category 1) include the work of Feichtenhofer et al. [80], that proposes a tracker based on R-FCN [81], an object detection framework which is fully convolutional up to region classification and regression and that is extended for multi-frame detection and tracking. The model is evaluated on ImageNet object detection from a video dataset [82] achieving an mAP (mean average precision) of 68.8% on bicycles and 79.8% on motorcycles. Meanwhile, Gunawan and Jatmiko [83] show an MOT with deep network embedding (category 2) proposing a Geometric Deep Particle Filter (GDPF) for motorcycle tracking, looking to improve the tracker's transition model and including a stacked denoising autoenconder (SDAE) [84] as a deep learning observation model. In spite of the success of the tracker in other domains, it has poor performance for motorcycles in an ad-hoc video dataset, mainly due to the vehicles' hard maneuvering. Overall, although deep learning methods show good promise, much remains to be done for tracking of motorcycles, especially in cluttered conditions.

VI. Datasets and Performance Measures

There is no clear consensus on the metrics or datasets to be used for research on motorcycle detection/tracking. This makes it difficult to compare results and to evaluate with fair criteria the different strategies used for detection and/or tracking.

A. Datasets

Realistic (manually annotated, long, varied traffic and weather conditions, etcds.) public datasets are still a necessity to assess and compare detection and tracking algorithms proposed by researchers. Generally, they have tended to develop their own private datasets making comparisons difficult, especially when people do not share algorithms. Clearly, public annotated data is very useful for training.

Nevertheless, there have been some efforts to produce useful public datasets. Bileschi et al. [85] describe the CBCL

TABLE II
DEEP LEARNING APPLIED TO MOTORCYCLE DETECTION

Algorithms and Ref.	Preprocessing and/or Characteristics	Cluttering/ Oclussion	DL Strategy	Dataset	Performance
MS- CNN [70]	Multi-scale Object Proposal Network	All	Based on Fast-RCNN	KITTI [60]	Recall 0.84 easy Recall 0.75 moderate Recall 66.07 hard
Urban motorcycle detection [72]	Scaling + Non-max suppression	Medium	CNN	Proprietary	0.81 F1-Score
Helmet detection using DL [75]	GMM Background Subtraction + CNN for features and classification (motorcycles and Helmet)	Low	CNN Based on AlexNet [52]	Proprietary	Accuracy 0.99 IITH Helmet 1 Accuracy 0.91 IITH Helmet 2
Helmet violation by using DL [76]	HOG+SVM, CNN and Haar Cascade + Segmented OCR	Low	AlexNet + LeNet	Proprietary	Validation Accuracy 0.98 Test Accuracy 0.97
DCNN for vehicle classification [68]	Selective Search	Low	Selective Search + CNN	Proprietary	100 Precision 0.89 Recall
Motorcycle classification [74]	Pre-segmented images	Low	Feature extraction by AlexNet + SVM	Proprietary	100 Accuracy (motorcycles)
Vehicle detection classification [73]	GMM	Medium	Feature extraction by AlexNet + SVM	Proprietary	F1 0.76 (All Vehicles)
Helmet Detection [77]	GMM	Low	Faster R-CNN	Proprietary	mAP

StreetScenes database which corresponds to more than 8,000 images of 1280 x 960 pixels. 3,547 of them were labeled and include categories such as car, pedestrian, bicycle, motorcycles, building, tree, road, sky, sidewalk and store. The Penn-Fudan dataset is introduced in [86] which includes pedestrian, bike, human riding bike, umbrella and car object classes, taken from scenes around campus and urban streets and has the segmented ground truth of the different classes. There is also the highly cited PASCAL 2 Visual Object Classes Challenge 2012 (VOC2012) [87], recorded from a stationary vehicle and popular for bicycle detection. It includes 20 different classes, containing bicycles, cars and motorcycles. For 3-D detection algorithms, there is also 174 motorcycle instances with different poses (12-16). The train/val data has 11,530 images containing 27,450 ROI annotated objects and 6,929 segmentations. This dataset is generally used for training purposes, to train both for cars and bicycles. Some authors complement the set to reduce overfitting and to increase the generalization capabilities of the classifier as in [88]. Cho et al. [89] made available a public dataset for tracking purposes, consisting of 6 video sequences, intending to capture the ego motion perceived by a vehicle. CityScapes Dataset [90] comprises a set of stereo video sequences captured in streets from 50 different cities (mostly in Germany). 5,000 of these images have high quality pixel-level annotations while 20,000 additional images have coarse annotations. The images were recorded during several months (spring, summer, fall) during daytime and under good and medium weather conditions. Specially oriented to motorcycles detection and classification and in some cases tracking, most authors present results working with their own datasets that are seldom made public as in [20], [21], [18], [25], [34], [35], [13], [14], [28], [37], [39], [36], [42], [43], [44], [45], [49] and [53]. This is a significant problem to compare results.

For deep learning strategies, mainly used in CNNs, there is the Caltech 256 [91] dataset that contains 30,608 images distributed in 256 different categories with at least 798 images of motorcycles 261 x 154 pixels but only in lateral views, which is a significant issue since the set explicitly avoids object rotation and different angles of view. Other datasets already described and used in DL are [60], [90], [92], [93], [94] and [95], which all share the property of having larger amounts of annotated data for training, but are not related to urban environments scenarios.

Recent developments related to autonomous driving benchmarks that evaluate object detection, stereo vision and semantic/instance segmentation, include the large-scale 5D semantic benchmark (BLVD) [96], which defines three kind of "participants": vehicles, pedestrians and riders (cyclists and motorbikes) due to its dynamic moving. This interesting benchmark is constructed by a self-driving platform and includes 249,129 3D annotations. Because of the different nature of the views from vehicle-based cameras, it is not clear how useful this data could be for traditional CCTV views.

B. Performance Measures

For classification and tracking, there are different strategies to measure the performance obtained when evaluating algorithms. For this, it is important to differentiate between detection, classification and tracking measures.

1) Detection Measures: For sliding windows, used to detect objects, researchers have used False Positives Per Window (FPPW) versus 1-Recall (FalseNeg/TruePos+FalseNeg) to generate the metric Detection Error Tradeoff (DET) [97]. The x- and y-axes are non-linearly scaled (using standard normal deviates or just by logarithmic transformation), which generates curves that are more linear than ROC curves and exploit most of the image area to highlight the differences. Also for detection, Bileschi et al. [85] use crop-wise detection measure, pixel-wise detection and box-wise detection measure, where an A object is considered to match with a baseline *B* object (ground truth) if area $A \cap B / A \cup B > \theta$, θ being a parameter that evaluates how close *A* and *B* should match. The default value is $\theta = 1/2$. This measure is also known as Intersection over Union (IoU) or Jaccard coefficient. Taking into account the above overlapping criteria, Everingham et al.

[98] propose mAP (mean Average Precision) for the Pascal Visual Object Classes Challenge (VOC).

$$mAP = \frac{1}{|Q_R|} \sum_{q \in Q_R} AP(q) \tag{1}$$

where Q_R is the set of image queries and q is the specific query. It is computed by averaging the precision values on the precision-recall curve where the recall is in the range $[0, 0.1, ..., 1]$ (e.g. average of 11 precision values). To be more precise, they consider a slightly corrected PR curve, where for each curve point (p, r), if there is a different curve point (p', r') such that $p' > p$ and $r' >= r$, it replaces p with maximum p' of those points.

For object counting, the Grid Average Mean absolute Error (GAME) [93] simultaneously considers the object count and the location estimated for the objects. It subdivides the image into 4 non-overlapping regions and computes the MAE in each of these sub regions.

$$GAME(L) = \frac{1}{N} \cdot \sum_{n=1}^{N} (\sum_{l=1}^{4} |e_n^l - gt_n^l|) \tag{2}$$

2) Classification Measures: Classification algorithms are typically evaluated through a confusion matrix [99] where, for unbalanced classes, it is better to use the F1 score, that could be extended as explained by Chen et al. [45] to deal with multiples classes. Metrics normally used for classification evaluation are normalized: recall *(REC)*, precision *(PRE)* and *(F1)*:

$$PRE = \frac{TP}{TP + FP} \tag{3}$$

$$REC = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \cdot REC \cdot PRE}{REC + PRE} \tag{5}$$

3) Tracking Measures: For single tracking algorithms Wu et al. [100], propose *OPE (One-time Pass Evaluation)*, which is initialized from the ground truth position of the first frame and the average precision or success rate is evaluated. Since the algorithm could be sensitive to initialization in the first frame and most algorithms do not have re-initialization mechanisms, the authors propose two other metrics: *Temporary robustness evaluation (TRE)*, where the evaluation starts from different initialization frame and *Spatial Robustness Evaluation (SRE)*, where different objects states with different shift or scaling of the ground truth are generated.

Dealing with Multiple Objects Tracking (MOT) Bernardin and Stiefelhagen [101] propose the *CLEAR MOT metrics*:
The *multiple object tracking precision (MOTP):*

$$MOTP = \frac{\sum i, td_t^i}{\sum t, C_t} \tag{6}$$

corresponds to the total error in matched object-hypothesis pairs over all frames, averaged over the total number of matches found.
The *multiple object tracking accuracy (MOTA):*

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{7}$$

where m_t, fp_t and mme_t are the number of misses, false positives and mismatches, respectively, for time t.

Later, Wen et al. [95] explain the importance of considering object detection and tracking jointly in MOT evaluation. For instance, they propose *PR-MOTA* which is a three-dimensional curve characterizing the relation between object detection performance (precision and recall) and object tracking performance (MOTA).

$$\Omega* = \frac{1}{2} \int_c \Psi(p, r) ds \tag{8}$$

where *c* is the *PR* curve and $\Psi(p, r)$ is the *MOTA* value corresponding to precision *p* and recall *r* on the *PR* curve. It is noted that although (or because) there is a rich set of metrics, different researchers use different sets, making it complicated to compare works.

VII. A BASELINE FOR FUTURE RESEARCHERS

A. The Urban Motorbike Dataset

Due to the lack of a urban motorcycle dataset, to serve as a common base for research in detection and tracking, we have created the public Urban Motorbike Dataset (UMD) [1] that contains images taken with a Phantom 4® drone, with an HD camera under windy conditions, which affected the image stabilizer. Images were resized to 640 x 364 pixels, containing 318 motorcycle tracks and 56,975 ROI annotated objects. 60% of the annotated data correspond to occluded motorcycles. Objects with heights less than 25 pixels were not annotated (See figure 4).

B. Preliminary evaluation

1) Detection: An improved version of the model proposed in [102] called EspiNet, is compared with some of the most representative models of deep learning single-stage detectors (Yolo V.3 [103]) and Region based detectors (Faster R-CNN [104]–(VGG16 based)). All these models were trained from scratch using the UMD dataset. According to recommended practice in deep learning, all three models use 90% of the data for training and 10% for validation. Training and test sets were taken randomly, to avoid bias.
Results shows that the EspiNet model achieves an Average Precision (AP) of 88.8% and an F1-score of 91.8%, outperforming results for YOLO and Faster R-CNN. Table III shows the evaluation results.

[1] http://videodatasets.org/UrbanMotorbike

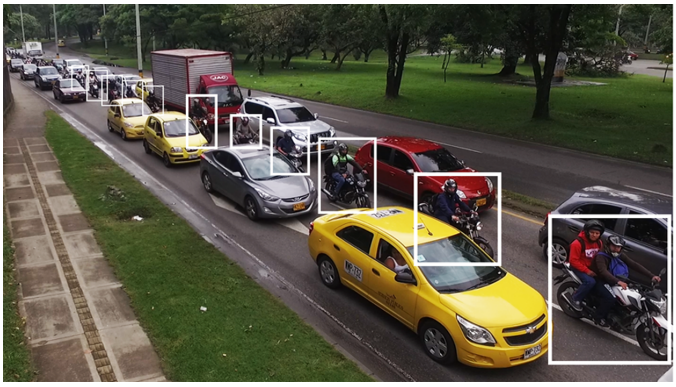

Fig. 4. UMD dataset - Note the significant level of occlusion between annotated motorcycles

TABLE III
COMPARATIVE DETECTION RESULTS FOR ESPINET, FASTER-RCNN (BASED ON VGG16) [104] AND YOLO V3 [103]

Metrics	EspiNet	Faster R-CNN	YOLO
Precision (%)	**93.7**	57.3	93.0
Recall (%)	**90.0**	76.3	81.0
F1 score (%)	**91.8**	65.4	86.6
AP (Average Precision)	**88.84**	68.75	80.75

2) Tracking: Two Multiple Object tracking methods are evaluated on the UMD dataset. In the first instance, an MOT framework based on Markov decision process (MDP) [105] is implemented. The tracker modeled the life time of the tracked object using four sub-space states (Active, Tracked, Lost and Inactive). The original algorithm has been evaluating with EspiNet, Yolo and Faster R-CNN (VGG16 based) as detectors. Additionally to Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA) (see section VI-B), another set of metrics are used [106] that allow understanding tracking behaviour using different detectors:

- Mostly Track targets (MT), percentage of ground truth tracks covered by the tracking mechanism for at least 80%.
- Mostly Lost targets (ML, percentage of ground truth tracks covered by the tracking mechanism less than 20%.
- False Positives (FP)
- False Negatives (FN)
- ID Switches (IDS) ID of the tracks that are erroneously changed by the algorithm.
- Fragmentations (Frag) the total number of times a trajectory is Fragmented.

Comparative results of MDP tracker applied to UMD dataset can be compared in table IV. The second MOT tracker used is a deep learning based tracker: DeepSort [107]. This tracker relies on motion and appearance information of the detected object to build the different tracks associated with an object. The tracking evaluation on the UMD dataset shows very close results to the obtained by using MDP tracker (see table V), illustrating that the good detection results obtained with EspiNet yield improved tracking results on an MOT tracker that uses tracking by detection as its main strategy.

TABLE IV
COMPARATIVE RESULTS FOR MDP TRACKING ON UMD DATASET, USING DETECTORS ESPINET [102], FASTER-RCNN (BASED ON VGG16) [64] AND YOLO [103].

Metrics	EspiNet	Faster R-CNN	YOLO
Recall	**92.5**	73.4	69.6
Precision	93.6	68.7	**95.7**
F1-Score	**93.0**	70.9	80.6
False Alarm Rate	0.33	1.76	**0.17**
GT Tracks	318	318	318
Mostly Tracked	**285**	107	69
Mostly Lost	**1**	7	7
False Positives	3,318	17,578	**1,661**
False Negatives	**3,922**	13,951	15,996
ID Swiches	**75**	662	80
Fragmentations	415	1,630	**299**
MOTA	**86.1**	38.7	66.2
MOTP	**77.7**	72.6	76.7

TABLE V
COMPARATIVE RESULTS FOR DEEPSORT TRACKER ON UMD DATASET, USING DETECTORS ESPINET [102], FASTER-RCNN (BASED ON VGG16) [64] AND YOLO [103].

Metrics	EspiNet	Faster R-CNN	YOLO
Recall	**91.1**	77.3	78.9
Precision	**96.6**	58.1	93.3
F1-Score	**93.7**	66.3	85.5
False Alarm Rate	**0.17**	2.93	0.30
GT Tracks	318	318	318
Mostly Tracked	**273**	140	159
Mostly Lost	5	6	**1**
False Positives	**1,704**	29,255	2,959
False Negatives	**4,698**	11,904	11,083
ID Swiches	**112**	1,958	286
Fragmentations	**555**	2,084	635
MOTA	**87.6**	17.9	72.7
MOTP	**77.2**	71.6	76.1

VIII. DISCUSSION

A. Challenges

The research topic of motorcycles detection and tracking, especially in urban environments, is still quite open due to different aspects that need to be considered to perform such tasks. Deep learning (DL) feature representation provides a powerful alternative over traditional features such as LBP, SURF, HOG and has shown good results in the standard object recognition challenges [62], [52], [51]. For motorcycles detection, important advances are found in the literature on feature-classifier-based detection. However, many of the articles discussed here are not yet suitable for real-time video surveillance scenarios, given the sliding windows strategy, which is time-consuming. In some implementations, GPU-based techniques are used to accelerate the computation, nevertheless it is necessary to find strategies that take advantage of the perspective of the scene (e.g. camera calibration), allowing to establish ROIs, reducing the analyzed region which increases the performance of the

proposed algorithms. Meanwhile DL for vehicle detection, exploits parallel architectures and, in successful implementations, it performs both region proposals and object detection [64], [66], [108], [109], [110], with near real-time operation.

Motorcycle hypothesis generation faces the challenge of having to deal with strong changes in appearance depending on camera viewpoints. Detectors based on deep learning architectures have shown to be the best option to store complex feature descriptors as the convolutional layers aggregates complexity in deeper layers. These DL detectors are successful mainly for on-board autonomous systems but are yet to be found in CCTV-based urban surveillance.

The literature reports few studies regarding motorcycle detection, given the complexity of the analyzed objects. Some of this complexity has to deal with phenomena such as shadows [34], [111], but there are also the problems of night time, poor illumination conditions and bad weather. Even though appearance features have been successfully applied to other types of vehicles, for motorcyclists the problem increases due to the small size area to detect, making it difficult to measure symmetry properties and to establish an adequate ROI. The same happens with color, corner, edges and texture features, which are affected by illuminations conditions and distance to the camera. Some of the algorithms use features extracted from the rider's helmet establishing relationships between vehicle and rider [39], but failing to consider real-world situations, for instance when the vehicle carries more than one passenger or when riders do not wear helmet protection (but see the work of Silva et al. [13], [14])

There is still very limited literature dealing with all ranges of weather and illuminations conditions. Generally, the research is framed to specific regulations and geographical regions.

The main technical challenge in motorcycles detection is related to real-world scenarios of traffic under congestion where occlusions could be frequent and using established CCTV infrastructure. In motorcycle detection only 10% of the analyzed algorithms work with some level of occlusion. Besides the strategies described above, there are some uses of 3-D models in scenarios where it is necessary to perform classification of vehicles in urban areas, which demands a previous camera calibration step for better results [28], [29], [30], [112]. There is some research [113] that uses DL working with 3-D models for vehicle detection, improving the results of previous point cloud based detection approaches. Some of the DL algorithms for motorcycle detection reviewed here are able to work in partly occluded scenarios, outperforming traditional methods. There are strategies that fuse RGB images and LiDAR points clouds as inputs for vehicle detection with high accuracy [114]. Building this type of datasets and using results derived from them requires access to LiDAR devices that might not be widely available, specially when considering existing CCTV infrastructure.

Comparing the different algorithms proposed in the literature is also a very difficult challenge, mainly due to the lack of universal ground truths to perform a fair benchmark. Most of the training sets of the algorithms described so far are not representative enough for the different urban contexts, unlike efforts for other type of road users such as the KITTI dataset [60] and Tsinghua-Daimler Cyclist Benchmark [12] for bicycles, Cars dataset [92], VeRi-776 dataset [115], TRANCOS [93], CompCars dataset [94] and CityScapes Dataset [90] for other vehicles. To palliate this we have introduced the UMD dataset (VII-A) specifically aimed for detection and tracking of motorcycles in urban environments and made available publicly to the research to advance the state-of-the-art in part of this field.

IX. Conclusions and Future Work

One interesting finding is that there exists relatively little literature on motorbikes detection and tracking. The different works are done in relatively simple environments and the algorithms reported are notoriously difficult to apply for data different from their original reports and in particular to be deployed for existing CCTV infrastructure.

Future research in motorcycle detection, would look to deal with the wide variability of real world images, including changes in illumination, drastic object scale changes and view point, deformations of parts, noisy images, blur resolution, poor weather conditions, day/night operation, etc. This will require time-intensive procurement of representative datasets and their corresponding ground truths.

Motorcycling detection needs to consider that in many places helmets are not worn, making drivers even more vulnerable. So their study it is also matter of public health and safety conditions.

The different works cited in this paper point to the need to improve motorcycles tracking strategies in realistic congested scenarios. Although motion-based methods are useful for tracking moving objects, they demand important computational resources and involve analysis of various previous frames for an object can be detected. They may fail to detect objects with slow relative motion and are also sensitive to camera movement. It is also worth to mention than despite the latest developments in Multiple Object Tracking (MOT) extensively reviewed in [79], to date there are virtually no studies of such methods for tracking motorcycles.

Deep Learning theory has demonstrated to be useful in the field of vehicle detection, but there is yet little work on DL applied to the detection of motorcycles. The principal advantage of this approach (applied to object detection) lies in the ability to learn richer invariant features via multiple non-linear transformations. Nevertheless, an important drawback of CNN appears in the convolutional layer at each trainable stage, the kernels/weights employed in the convolution are trained by back propagation algorithm, which is time- and data-consuming. Several strategies as *pretraining* have been proposed to overcome this issue, which becomes a prerequisite to implement real-time applications based on CNN.

Region based detectors are popular, covering region proposals and detection modules, that have evolved from R-CNN, Fast R-CNN and Faster R-CNN reaching even pixel level segmentation in Mask R-CNN [116]. However, these architectures demand large amounts of examples to be able to achieve acceptable results as in all DL models. Most promising results in real time DL are being obtained by single stage

detectors such as SSD [69], YOLO [117] and specially RRC [118], which is a state of the art detector, even though object detection may fail where objects are really small or may appears quite close each other in the scene.

One interesting topic for deep learning detectors is to incorporate context reasoning. Despite that before the irruption of DL, contextual information was already regarded as useful for improving detection and recognition algorithms, it has not been extensively studied in the context of deep learning. This topic offer many possibilities for research, more in the context-rich urban environments.

An important aspect in motorcycle detection is the need to have realistic public datasets to allow researchers to benchmark different algorithms in many scenarios, validating the results in a structured and common way.

Most of the promising employed motorcycle detectors rely on fully supervised learning schemes. This requires annotated data (ground truth), something that is expensive specially for DL approaches. It would be interesting to start studying semi supervised or unsupervised models for motorcycle and bicycle detectors.

ACKNOWLEDGMENTS

S.A. Velastin is grateful to funding received from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 600371, el Ministerio de Economía y Competitividad (COFUND2013-51509) and Banco Santander. This work was partially supported by COLCIENCIAS project: Reduccion de Emisiones Vehiculares Mediante el Modelado y Gestion Optima de Trafico e n A reas M etropolitanas - C aso M edellin - Area Metropolitana del Valle de Aburra, codigo 111874558167, CT 049-2017. Universidad Nacional de Colombia. Proyecto HERMES 25374. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

REFERENCES

[1] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[2] "Urban Transport - Regions," [Online; accessed 2016-10-26]. [Online]. Available: https://goo.gl/MwWrsG

[3] "Motociclistas, un grave problema vial en Colombia," [Online; accessed 2016-10-26]. [Online]. Available: http://www.dicyt.com/noticias/motociclistas-un-grave-problema-vial-en-colombia

[4] H. S. Bazargani, R. G. Vahidi, and A. A. Abhari, "Predictors of Survival in Motor Vehicle Accidents Among Motorcyclists, Bicyclists and Pedestrians," *Trauma Monthly*, vol. 22, no. 2, 2017, [Online; accessed 2017-09-26]. [Online]. Available: http://traumamon.neoscriber.org/en/articles/13364.html

[5] E. P. S.A, "El 50% de las motos que circulan en Cali son 'foráneas'," [Online; accessed 2016-11-26]. [Online]. Available: http://www.elpais.com.co/elpais/cali/noticias/50-motos-circulan-cali-son-foraneas

[6] "WHO | Global status report on road safety 2018." [Online]. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

[7] A. Constant and E. Lagarde, "Protecting vulnerable road users from injury," *PLoS Med*, vol. 7, no. 3, p. e1000228, 2010.

[8] H. Cho, P. E. Rybski, and W. Zhang, "Vision-based bicycle detection and tracking using a deformable part model and an EKF algorithm." 2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), 9 2010, pp. 1875–1880.

[9] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 7 2010.

[10] M. Enzweiler and D. Gavrila, "Monocular Pedestrian Detection: Survey and Experiments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 12 2009.

[11] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten Years of Pedestrian Detection, What Have We Learned?" *arXiv:1411.4304 [cs]*, 11 2014, arXiv: 1411.4304. [Online]. Available: http://arxiv.org/abs/1411.4304

[12] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, "A new benchmark for vision-based cyclist detection." 2016 IEEE Intelligent Vehicles Symposium (IV), 6 2016, pp. 1028–1033.

[13] R. R. e Silva, K. R. Aires, and R. de MS Veras, "Detection of helmets on motorcyclists," *Multimedia Tools and Applications*, p. 1–25, 2017.

[14] ——, "Detection of helmets on motorcyclists," *Multimedia Tools and Applications*, vol. 77, no. 5, pp. 5659–5683, 2018.

[15] A. Mukhtar and T. B. Tang, "Vision based motorcycle detection using HOG features." 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 10 2015, pp. 452–456, cites=?

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, p. 91–110, 2004.

[17] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms." ACM, 2010, p. 1469–1472, [Online; accessed 2016-02-12]. [Online]. Available: http://dl.acm.org/citation.cfm?id=1874249

[18] N. D. Thai, T. S. Le, N. Thoai, and K. Hamamoto, "Learning bag of visual words for motorbike detection." 2014 13th International Conference on Control Automation Robotics Vision (ICARCV), 12 2014, pp. 1045–1050.

[19] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel." ACM, 2007, p. 401–408, [Online; accessed 2016-01-19]. [Online]. Available: http://dl.acm.org/citation.cfm?id=1282340

[20] K. Dahiya, D. Singh, and C. K. Mohan, "Automatic detection of bike-riders without helmet using surveillance videos in real-time." 2016 International Joint Conference on Neural Networks (IJCNN), 7 2016, pp. 3046–3051.

[21] D. Singh, C. Vishnu, and C. K. Mohan, "Visual Big Data Analytics for Traffic Monitoring in Smart City." 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 12 2016, pp. 886–891.

[22] S. A. Ghonge and J. B. Sanghavi, "Smart Surveillance System for Automatic Detection of License Plate Number of Motorcyclists without Helmet," 2018.

[23] A. S. Talaulikar, S. Sanathanan, and C. N. Modi, "An Enhanced Approach for Detecting Helmet on Motorcyclists Using Image Processing and Machine Learning Techniques," in *Advanced Computing and Communication Technologies*. Springer, 2019, pp. 109–119.

[24] Barış and Y. Baştanlar, "Classification and tracking of traffic scene objects with hybrid camera systems." Institute of Electrical and Electronics Engineers, 2018.

[25] Y. Shuo and E.-J. Choi, "A Driving Support System Base on Traffic Environment Analysis," *Indian Journal of Science and Technology*, vol. 9, no. 47, 12 2016, [Online; accessed 2017-09-26]. [Online]. Available: http://www.indjst.org/index.php/indjst/article/view/108374

[26] P. Wonghabut, J. Kumphong, T. Satiennam, R. Ung-arunyawee, and W. Leelapatra, "Automatic helmet-wearing detection for law enforcement using CCTV cameras," in *IOP Conference Series: Earth and Environmental Science*, vol. 143. IOP Publishing, 2018, p. 012063.

[27] R. J. Gavadi and S. S. Patil, "Automatic Detection of Motorcyclist without Helmet using Haar Cascade Classifier," *Journal of Integrated Science and Technology*, vol. 6, no. 2, pp. 33–36, 2018.

[28] S. Messelodi, C. M. Modena, and G. Cattoni, "Vision-based bicycle/motorcycle classification," *Pattern recognition letters*, vol. 28, no. 13, p. 1719–1726, 2007.

[29] S. Messelodi, C. M. Modena, and M. Zanin, "A computer vision system for the detection and classification of vehicles at urban road intersections," *Pattern analysis and applications*, vol. 8, no. 1-2, p. 17–31, 2005.

[30] N. Buch, J. Orwell, and S. A. Velastin, "Urban road user detection and classification using 3D wire frame models," *IET Computer Vision*, vol. 4, no. 2, pp. 105–116, 6 2010.

[31] T. S. Le and C. K. Huynh, "An Unified Framework for Motorbike Counting and Detecting in Traffic Videos." 2015 International Conference on Advanced Computing and Applications (ACOMP), 11 2015, pp. 162–168.

[32] M. Muzammel, M. Z. Yusoff, and F. Meriaudeau, "Rear-end vision-based collision detection system for motorcyclists," *Journal of Electronic Imaging*, vol. 26, no. 3, p. 033002, 5 2017.

[33] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," *a talk at the Stanford Artificial Project in*, p. 271–272, 1968.

[34] B. Duan, W. Liu, P. Fu, C. Yang, X. Wen, and H. Yuan, "Real-time on-road vehicle and motorcycle detection using a single camera." IEEE, 2009, p. 1–6, [Online; accessed 2016-01-20]. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4939585

[35] M. A. M. Nong, R. Osman, J. M. Yusof, and R. M. Sidek, "Motorcycle image tracking and edge detections based on Simulink software." 2016 6th International Conference on Intelligent and Advanced Systems (ICIAS), 8 2016, pp. 1–4.

[36] N. Kanhere, S. Birchfield, W. Sarasua, and S. Khoeini, "Traffic monitoring of motorcycles during special events using video detection," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2160, p. 69–76, 2010.

[37] C.-C. Chiu, M.-Y. Ku, and H.-T. Chen, "Motorcycle detection and tracking system with occlusion segmentation." IEEE, 2007, p. 32–32, [Online; accessed 2016-01-20]. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4279140

[38] M.-Y. Ku, C.-C. Chiu, H.-T. Chen, and S.-H. Hong, "Visual motorcycle detection and tracking algorithms," *WSEAS Transaction on electronics*, p. 121–131, 2008.

[39] J. Chiverton, "Helmet presence classification with motorcycle detection and tracking," *Intelligent Transport Systems, IET*, vol. 6, no. 3, p. 259–269, 2012.

[40] S. Sekar, S. Kulam, S. Selvaraj, and P. Saravanan, "An Automatic Helmet Detection and Penalty System Using Image Descriptors and Classifiers," *Journal of Computational and Theoretical Nanoscience*, vol. 15, no. 6-7, pp. 2245–2250, 2018.

[41] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," vol. 2. Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., 1999, pp. –252 Vol. 2.

[42] R. Waranusast, N. Bundon, V. Timtong, C. Tangnoi, and P. Pattanathaburt, "Machine vision techniques for motorcycle safety helmet detection." IEEE, 2013, p. 35–40, cites=8. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6726989

[43] Y. Dupuis, P. Subirats, and P. Vasseur, "Robust image segmentation for overhead real time motorbike counting." 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), 10 2014, pp. 3070–3075.

[44] M. A. Rashidan, Y. M. Mustafah, A. A. Shafie, N. A. Zainuddin, N. N. A. Aziz, and A. W. Azman, "Moving Object Detection and Classification Using Neuro-Fuzzy Approach," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 11, no. 4, p. 253–266, 2016.

[45] Z. Chen, T. Ellis, and S. A. Velastin, "Vehicle detection, tracking and classification in urban traffic." 2012 15th International IEEE Conference on Intelligent Transportation Systems, 9 2012, pp. 951–956, cites=37.

[46] Z. Chen and T. Ellis, "Self-adaptive Gaussian mixture model for urban traffic monitoring system." 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 11 2011, pp. 1769–1776.

[47] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers." ACM, 1992, p. 144–152, [Online; accessed 2015-03-21]. [Online]. Available: http://dl.acm.org/citation.cfm?id=130401

[48] A. González, D. Vázquez, A. M. López, and J. Amores, "On-Board Object Detection: Multicue, Multimodal, and Multiview Random Forest of Local Experts," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–11, 2017.

[49] S. Sutikno, I. Waspada, N. Bahtiar, and P. S. Sasongko, "Classification of Motorcyclists not Wear Helmet on Digital Image with Backpropagation Neural Network," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 14, no. 3, p. 11281133, 9 2016.

[50] S. Ojha and S. Sakhare, "Image processing techniques for object tracking in video surveillance- A survey." 2015 International Conference on Pervasive Computing (ICPC), 1 2015, pp. 1–6.

[51] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks." Springer, 2014, p. 818–833, [Online; accessed 2016-09-07]. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-10590-1_53

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," 2012, p. 1097–1105, [Online; accessed 2016-09-05]. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-w

[53] Z. Chen and T. Ellis, "Multi-shape Descriptor Vehicle Classification for Urban Traffic." 2011 International Conference on Digital Image Computing Techniques and Applications (DICTA), 12 2011, pp. 456–461.

[54] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, 12 2006, [Online; accessed 2013-04-10]. [Online]. Available: http://doi.acm.org/10.1145/1177352.1177355

[55] M. H. Zaki, T. Sayed, and X. Wang, "Computer vision approach for the classification of bike type (motorized versus non-motorized) during busy traffic in the city of Shanghai," *Journal of Advanced Transportation*, vol. 50, no. 3, pp. 348–362, 4 2016.

[56] C. Harris and M. Stephens, "A combined corner and edge detector." vol. 15. Manchester, UK, 1988, p. 50, [Online; accessed 2015-03-18]. [Online]. Available: http://courses.daiict.ac.in/pluginfile.php/13002/mod_resource/content/0/References/harris1988.pdf

[57] M. MOHAMED and N. Saunier, "Applications of Multi-Level Pattern Learning to Traffic Scene Interpretation and Anomaly Detection," Tech. Rep., 2017.

[58] S. Jackson, L. Miranda-Moreno, P. St-Aubin, and N. Saunier, "Flexible, Mobile Video Camera System and Open Source Video Analysis Software for Road Safety and Behavioral Analysis," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2365, pp. 90–98, 12 2013.

[59] W. Liu, X. Wen, B. Duan, H. Yuan, and N. Wang, "Rear Vehicle Detection and Tracking for Lane Change Assist." 2007 IEEE Intelligent Vehicles Symposium, 6 2007, pp. 252–257.

[60] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite." IEEE, 2012, p. 3354–3361, [Online; accessed 2016-10-27]. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6248074

[61] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]," *IEEE Computational Intelligence Magazine*, vol. 5, no. 4, pp. 13–18, 11 2010.

[62] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database." IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, 6 2009, pp. 248–255.

[63] P. Soviany and R. T. Ionescu, "Optimizing the Trade-off between Single-Stage and Two-Stage Object Detectors using Image Difficulty Prediction," *arXiv:1803.08707 [cs]*, Mar. 2018, arXiv: 1803.08707. [Online]. Available: http://arxiv.org/abs/1803.08707

[64] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015, p. 91–99. [Online]. Available: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

[65] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." 2014 IEEE Conference on Computer Vision and Pattern Recognition, 6 2014, pp. 580–587.

[66] R. Girshick, "Fast r-cnn," 2015, p. 1440–1448. [Online]. Available: http://www.cv-foundation.org/openaccess/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html

[67] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, p. 154–171, 2013.

[68] Y. O. Adu-Gyamfi, S. K. Asare, A. Sharma, and T. Titus, "Automated Vehicle Recognition with Deep Convolutional Neural Networks," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2645, p. 113–122, 2017.

[69] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," *arXiv:1512.02325 [cs]*, vol. 9905, pp. 21–37, 2016, arXiv: 1512.02325.

[70] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014, [Online; accessed 2016-09-05]. [Online]. Available: http://arxiv.org/abs/1409.1556

[71] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 9 2010.

[72] C. K. Huynh, T. S. Le, and K. Hamamoto, "Convolutional neural network for motorbike detection in dense traffic." 2016 IEEE Sixth International Conference on Communications and Electronics (ICCE), 7 2016, pp. 369–374.

[73] J. E. Espinosa, S. A. Velastin, and J. W. Branch, "Vehicle Detection Using Alex Net and Faster R-CNN Deep Learning Models: A Comparative Study," in *International Visual Informatics Conference*. Springer, 2017, pp. 3–15.

[74] ——, "Motorcycle classification in urban scenarios using convolutional neural networks for feature extraction," in *8th International Conference of Pattern Recognition Systems (ICPRS 2017)*, Jul. 2017, pp. 1–6.

[75] C. Vishnu, D. Singh, C. K. Mohan, and S. Babu, "Detection of motorcyclists without helmet in videos using convolutional neural network." 2017 International Joint Conference on Neural Networks (IJCNN), 5 2017, pp. 3036–3041.

[76] K. C. D. Raj, A. Chairat, V. Timtong, M. N. Dailey, and M. Ekpanyapong, "Helmet violation processing using deep learning." 2018 International Workshop on Advanced Image Technology (IWAIT), 1 2018, pp. 1–4.

[77] B. Yogameena, K. Menaka, and S. S. Perumaal, "Deep learning-based helmet wear analysis of a motorcycle rider for intelligent surveillance system," *IET Intelligent Transport Systems*, 2019.

[78] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognition*, vol. 76, p. 323–338, 2018.

[79] S. Chen, Y. Xu, X. Zhou, and F. Li, "Deep Learning for Multiple Object Tracking: A Survey," *IET Computer Vision*, Jan. 2019. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-cvi.2018.5598

[80] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to Track and Track to Detect," 10 2017, [Online; accessed 2018-02-28]. [Online]. Available: http://search.arxiv.org:8081/paper.jsp?r=1710.03958&qid=1519851277477swap_nCnN_-641548753&qs=Detect+to+Track+and+Track+to+Detec&in=cs

[81] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," p. 9.

[82] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, p. 211–252, 2015.

[83] A. A S Gunawan and W. Jatmiko, "GEOMETRIC DEEP PARTICLE FILTER FOR MOTORCYCLE TRACKING: DEVELOPMENT OF INTELLIGENT TRAFFIC SYSTEM IN JAKARTA," *International Journal on Smart Sensing and Intelligent Systems*, vol. 8, no. 1, pp. 429–463, 2015.

[84] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," 2013, p. 809–817.

[85] S. M. Bileschi, "StreetScenes: Towards scene understanding in still images," Ph.D. dissertation, 2006, [Online; accessed 2016-10-28]. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.3289&rep=rep1&type=pdf

[86] L. Wang, J. Shi, G. Song, and I.-f. Shen, "Object detection combining recognition and segmentation." Springer, 2007, p. 189–199, [Online; accessed 2016-11-09]. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-76386-4_17

[87] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[88] A. Chayeb, N. Ouadah, Z. Tobal, M. Lakrouf, and O. Azouaoui, "HOG based multi-object detection for urban navigation." 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 10 2014, pp. 2962–2967.

[89] H. Cho, P. E. Rybski, and W. Zhang, "Vision-based 3D bicycle tracking using deformable part model and Interacting Multiple Model filter." 2011 IEEE International Conference on Robotics and Automation (ICRA), 5 2011, pp. 4391–4398.

[90] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6 2016, pp. 3213–3223.

[91] "Caltech256," [Online; accessed 2016-11-01]. [Online]. Available: http://vision.caltech.edu/Image_Datasets/Caltech256/#References

[92] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," 2013, p. 554–561. [Online]. Available: http://www.cv-foundation.org/openaccess/content_iccv_workshops_2013/W19/html/Krause_3D_Object_Representations_2013_ICCV_paper.html

[93] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Oñoro-Rubio, "Extremely overlapping vehicle counting." Springer, 2015, p. 423–431. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-19390-8_48

[94] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," 2015, p. 3973–3981. [Online]. Available: http://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Yang_A_Large-Scale_Car_2015_CVPR_paper.html

[95] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *arXiv preprint arXiv:1511.04136*, 2015. [Online]. Available: https://arxiv.org/abs/1511.04136

[96] J. Xue, J. Fang, T. Li, B. Zhang, P. Zhang, Z. Ye, and J. Dou, "Blvd: Building a large-scale 5d semantics benchmark for autonomous driving," *arXiv preprint arXiv:1903.06405*, 2019.

[97] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," Tech. Rep., 1997.

[98] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, p. 303–338, 2010.

[99] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment*, vol. 62, no. 1, p. 77–89, 1997.

[100] Y. Wu, J. Lim, and M.-H. Yang, "Online Object Tracking: A Benchmark." 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6 2013, pp. 2411–2418.

[101] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 1–10, 2008.

[102] J. E. Espinosa, S. A. Velastin, and J. W. Branch, "Motorcycle detection and classification in urban Scenarios using a model based on Faster R-CNN," *arXiv preprint arXiv:1808.02299*, 2018.

[103] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv:1804.02767 [cs]*, Apr. 2018, arXiv: 1804.02767. [Online]. Available: http://arxiv.org/abs/1804.02767

[104] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," *arXiv:1611.10012 [cs]*, Nov. 2016, arXiv: 1611.10012. [Online]. Available: http://arxiv.org/abs/1611.10012

[105] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4705–4713.

[106] F. Yin, D. Makris, and S. A. Velastin, "Performance evaluation of object tracking algorithms," 2007, p. 25. [Online]. Available: https://pdfs.semanticscholar.org/ad76/bdc7d06a7ec496ac788d667c6ad5fcc0fe41.pdf

[107] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," *arXiv:1703.07402 [cs]*, Mar. 2017, arXiv: 1703.07402. [Online]. Available: http://arxiv.org/abs/1703.07402

[108] Q. Fan, L. Brown, and J. Smith, "A closer look at Faster R-CNN for vehicle detection." 2016 IEEE Intelligent Vehicles Symposium (IV), 6 2016, pp. 124–129.

[109] S. Wang, F. Liu, Z. Gan, and Z. Cui, "Vehicle type classification via adaptive feature clustering for traffic surveillance video." 2016 8th International Conference on Wireless Communications Signal Processing (WCSP), 10 2016, pp. 1–5.

[110] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau, "Deep MANTA: A Coarse-to-fine Many-Task Network for joint 2D and 3D vehicle analysis from monocular image," *arXiv preprint arXiv:1703.07570*, 2017. [Online]. Available: https://arxiv.org/abs/1703.07570

[111] X. Wen, H. Yuan, C. Song, W. Liu, and H. Zhao, "An algorithm based on SVM ensembles for motorcycle recognition," 2007, p. 1–5, [Online; accessed 2016-01-20]. [Online]. Available: http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ieee-000004456371

[112] K. Schindler, A. Ess, B. Leibe, and L. Van Gool, "Automatic detection and tracking of pedestrians from a moving stereo rig," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 523–537, 11 2010.

[113] B. Li, "3D Fully Convolutional Network for Vehicle Detection in Point Cloud," *arXiv preprint arXiv:1611.08069*, 2016, [Online; accessed 2017-02-14]. [Online]. Available: https://arxiv.org/abs/1611.08069

[114] J. Dou, J. Xue, and J. Fang, "Seg-voxelnet for 3d vehicle detection from rgb and lidar data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4362–4368.
[115] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos." IEEE, 2016, p. 1–6. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7553002/
[116] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv:1703.06870 [cs]*, 3 2017, arXiv: 1703.06870. [Online]. Available: http://arxiv.org/abs/1703.06870
[117] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016, p. 779–788.
[118] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, "Accurate Single Stage Detector Using Recurrent Rolling Convolution," *arXiv:1704.05776 [cs]*, 4 2017, arXiv: 1704.05776. [Online]. Available: http://arxiv.org/abs/1704.05776

Jorge E. Espinosa was born in Bogotá D.C., Colombia, in 1973. He received the B.S. in System Engineering from the Universidad Los Libertadores de Colombia, in 2001 and M.Sc. degrees in Artificial Intelligence from the Katholieke Universiteit Leuven, Belgium, in 2003, For year 2019, he received a Meritorious Doctorate Thesis from the Universidad Nacional de Colombia.
From 2010, he has been full time professor at the Politécnico Colombiano Jaime Isaza Cadavid, Medellín Colombia. Teaching in areas as Programming and Artificial Intelligence. In the last 10 years as a professor, he has scientific publications in national and international journals and congresses, mostly related to his research theme: Optimization, Artificial Intelligence and related areas.

Sergio A. Velastín (M'90, SM'12) received the B.Sc. and M.Sc. (Research) degrees in electronics and the Ph.D. degree from the University of Manchester, Manchester, U.K., in 1978, 1979 and 1982, respectively, for research on vision systems for pedestrian tracking and road-traffic analysis. He worked in industrial R&D before joining King's College London (UK) in 1991 and then Kingston University London where he became director of its Digital Imaging Research Centre and full professor of applied computer vision. In 2015 he moved to the University Carlos III of Madrid, Spain where he was a Marie Curie Professor. He is a Fellow of the IET and currently Senior Research Scientist at Zebra Technologies Corp. and a visiting professor at Universidad Carlos III in Madrid and at Queen Mary University of London, UK.

John W. Branch He received the B.S. and M.S. degrees in Mines and Metallurgic Engineering from the Universidad Nacional de Colombia, Medellín, in 1995 and 1997 respectively and the Ph.D. degree in systems engineering from Universidad Nacional de Colombia, in 2007.
From 2000, he has been full time professor at the Universidad Nacional de Colombia. In the last 15 years as a professor, he has made 160 scientific publications in national and international journals and congresses, mostly related to his research theme: Computer Vision and related areas. He has directed 30 master's thesis and three doctoral thesis, currently directs ten doctoral thesis and five master's theses.
Prof. Branch has received the following prizes and distinctions, Thesis of Meritorious Degree. Postgraduate in Systems Engineering. Universidad Nacional de Colombia - Medellín – 1997. Meritorious Degree Work. Mining and Metallurgy Engineering, Universidad Nacional de Colombia - Medellín – 1995. Fourth Place Best Grade Contest VI version, Universidad Nacional de Colombia – 1996. Meritorious Doctorate Thesis, Universidad Nacional de Colombia - Medellin - 2007.
.