Neurocomputing 494 (2022) 43–55

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Survey paper

Multi-object tracking in traffic environments: A systematic literature review

Diego M. Jiménez-Bravo [a,*], Álvaro Lozano Murciego [a], André Sales Mendes [a], Héctor Sánchez San Blás [a], Javier Bajo [b]

[a] Expert Systems and Applications Lab, Department of Computer Science and Automatics, Faculty of Sciences, University of Salamanca, Salamanca 37008, Spain
[b] Ontology Engineering Group, Artificial Intelligence Department, Universidad Politécnica de Madrid, Madrid 28040, Spain

ARTICLE INFO

Article history:
Received 2 November 2021
Revised 6 April 2022
Accepted 17 April 2022
Available online 20 April 2022
Communicated by Zidong Wang

Keywords:
CNNs
Datasets
Evaluation metrics
MOT
Multi-object tracking
SLR
Systematic literature review
Traffic environments

ABSTRACT

The use of computer vision techniques to detect objects in images has grown in recent years. These techniques are especially useful to automatically extract and analyze information from an image or a sequence of them. One of the problems addressed by computer vision is multi-object tracking over frames sequences. To know the path and direction of objects can be crucial for some areas like traffic control and supervision; by doing that the system can be able to reduce traffic jams or redirect vehicles over less condensed areas. These algorithms include several aspects to have in mind in order to start a new development or research in this area, for instance, is important to review the current state-of-the art techniques, the hardware requirements, the main evaluation metrics, the commonly used datasets, among others. Therefore, the objective of this research is to present a systematic literature review which analyzes the recent works developed in the area of multi-object tracking in traffic environments. This paper reviews the techniques, hardware, datasets, metrics, and open lines of research in this area.

1. Introduction

During the last years, the field of computer vision research has suffered a big evolution and nowadays is one of the most active areas of computer science research. Numerous researches have emerged using and proposing novel architectures and methods in computer vision such as, R-CNN (Region-Convolutional Neural Network) [1], Fast R-CNN [2], Fater R-CNN [3], and YOLO (You Only Look Once) [4]. These novel methods are in many cases developed by big tech companies, for instance, Facebook [3,4], Microsoft [1], Google, etc. Thus, this underlines the fact of the importance of computer vision in the actual research overview. Many of these researches are applied in object detection and object tracking; both of them are extremely related since object tracking depends on object detection.

In a simple way, object tracking can be defined as the process of tracking an object through different frames in such a way that its position and direction throughout the sequence is known. In this way the tracking task can include two main sub-tasks, the detection of objects and the identification or re-identification of objects between the different frames. Thanks to these two tasks it is possible to obtain the exact position of the different objects in an image and the path they have followed in the field of view represented by the images.

These research areas are extremely useful since they help to monitor easily and rapidly open or crowded environments, for instance, shopping centers, public buildings, squares, streets, highways, etc. The task of monitoring several moving objects in a recording becomes more challenging for a human being as the number of objects to track grows [5]. On the contrary, computer vision can perform detection and tracking activities over hundreds of elements with excellent results. This is the field of multi-object tracking.

Nonetheless, it is not as easy as it sounds. On the one hand, detections techniques have to identify where an object is and what object it is. In order to do that, the techniques have to determine the background of the objects and deal with the different poses, colors, etc. On the other hand, tracking algorithms deal with more problems, for instance, identify new objects, reidentify lost objects, occlusion, background clutter, and pose changes.

* Corresponding author.
E-mail addresses: dmjimenez@usal.es (D.M. Jiménez-Bravo), loza@usal.es (Á. Lozano Murciego), andremendes@usal.es (A. Sales Mendes), hectorsanchezsanblas@usal.es (H. Sánchez San Blás), jbajo@fi.upm.es (J. Bajo).

https://doi.org/10.1016/j.neucom.2022.04.087
0925-2312/

As we briefly mentioned one of the areas where detection and tracking techniques are used is in the traffic environments. For many companies and especially city councils it is crucial to know how vehicles users operate on the road. This information is useful to detect dangerous activities on the road, to facilitate the traffic flow in an intersection or crowded roads, among other uses. Nevertheless, all the problems mentioned in the previous paragraph have to be taken into account. Moreover, the tracking system may have to consider the information and data obtained with other sensors or cameras located or not in the same place.

As we can see the systems that implement detection and tracking technologies are extremely useful for humanity and at the same time its complexity is enormous. However, there is no recent systematic review that addresses these systems in the domain of traffic environments to analyse the current state of the art and facilitate research in this field. Therefore, we consider that is important to review recent articles in the field of multi-object tracking applied to traffic environments to properly know the most used and novel techniques, the different types of datasets, the hardware used in the different studies, the evaluation metrics commonly used and the open lines of research of the shape of things to come. Consequently, we performed a SLR (Systematic Literature Review).

The rest of the article is structured as follows: Section 2 explains the methodology and the review planning implemented for this SLR; Section 3 answers the research questions of the SLR; in Section 4 we discuss the more relevant aspects of Section 3; and finally, Section 5 points out the conclusions and future lines of research.

2. Methodology and review process

For the development of this Systematic Literature Review we have followed the guidelines of Kitchenham and Charters in [6]. To guarantee that there are no other similar state-of-the-art reviews published in the recent past, we perform a search in the most important scientific articles databases. The search was performed in databases such as IEEE Xplore, Web of Science, SpringerLink, ACM Digital Library or Scopus within the terms "slr", "review", "survey", "state of the art", "mot", "multi-object tracking", "traffic", ... After performing the searches, we confirmed that there are not review concerning the research area of this article; hence, this outcome justifies the development of this work.

Therefore, and following the proposed methodology, we must start with the review process. In the following subsections we show the most relevant points or issues of the review process. It is important to note that not all of them are included, only those that we consider most important to illustrate the process.

2.1. Research questions

The research questions define the outcomes of the study; they establish what we want to survey with this Systematic Literature Review. The research questions of this study are related with MOT (Multi-Object Tracking) in traffic environments. The questions are as follows:

- **RQ1**: Which are the main techniques for Multi-Object Tracking in traffic environments?
- **RQ2**: Which are the devices used for Multi-Object Tracking in traffic environments?
- **RQ3**: Which are the main datasets for Multi-Object Tracking in traffic environments?
- **RQ4**: Which are the evaluation metrics for Multi-Object Tracking in traffic environments?
- **RQ5**: What are the main open lines of research or issues in this domain?

2.2. Search strategy

Another important issue to consider when designing the review process is the search strategy. The search strategy involves the databases that are going to be used as well as the concepts which are going to be used in the search queries. Hence, the concepts are going to form a query which is going to be consulted in every one of the databases selected for the study.

On the one hand, the current study makes use of four databases: IEEE Xplore, Web of Science, SpringerLink, and Scopus. These databases are commonly used to publish relevant articles in the topic's research area. Also, the advance query options are very similar between them; this helped us to form the queries.

On the other hand, the concepts used are the following ones:

- The term "multi-object tracking" or "multi object tracking" or "MOT" or "multi-target tracking" or "multi target tracking" to search for papers that study the multi-object tracking problem in computer vision.
- The term "traffic" to delimitate the study to those articles that develop a MOT technique in a traffic environment.
- The term "urban" or "environment" or "vehicle" or "city" or "road" or "lane" or "motorway" to refer to the context of "traffic" environments.

Once the whole process has been defined, the systematic review process can begin. This process is summarised in 1.

Next, it is possible to begin to answer the research questions mentioned previously.

3. Results

The area of computer vision research has different sub-areas that have or have had great relevance over the years. Perhaps the first major achievement of computer vision was image classification [7]. The classification of images consists of identifying the object or objects that appear in the image, in this way the output of the process will be the object that appears in the image or if it is a multi-label classification, the objects that appear in the image. However, image classification is only able to predict which elements are present in the image and not where they are, hence the existence of object detection techniques. These techniques detect the exact position of objects in an image and are capable of detecting an infinite number of object types as long as they have been considered during the training process.

Object detection techniques have been a great revolution in the area of computer vision and are one of the fundamental pillars used in image tracking techniques and therefore in MOT. MOT can be defined as the task of locating various entities or objects by associating an identifier with them and tracking their trajectory through the different frames of a given video or sequence of frames as input. Hence, in this section we are going to go deeper into the studies carried out in MOT by answering the questions mentioned in Section 2.1.

3.1. Which are the main techniques for Multi-Object Tracking in traffic environments?

To answer this question, we believe that is better to introduce the problem formulation of MOT tasks. MOT can be seen as a sequence of images in which algorithms obtain different observations of objects that have to match with the states of these objects

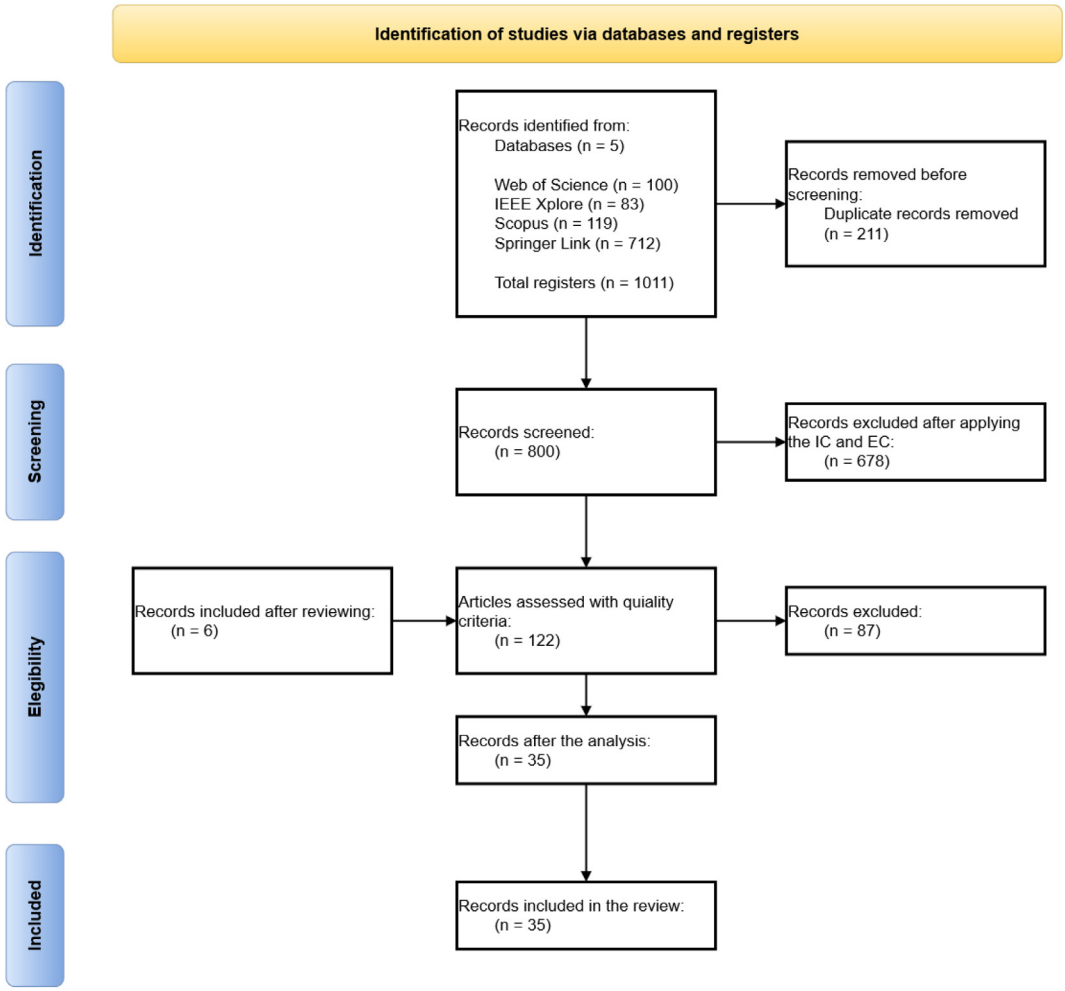

Fig. 1. Review process using the PRISMA 2020 flow diagram.

in the different frames of the sequence. In other words, the MOT algorithm has to obtain the different states (trace) of the objects in the different images but relating the same object over the different images.

In general terms, the MOT problem can be applied in two different situations. These situations depend on the number of cameras that we use for the study. Therefore, we can have MOT with a single-camera view or MOT with a multi-camera view. The second approach is more difficult than the first one because multi-camera view MOT has to deal with problems with overlapping and non-overlapping cameras, the tracking of objects over different views, and topology calculation among other problems. The selected papers of the study address a solution with a single-camera view.

However, it is not the only classification of MOT problems. For instance, depending on the initialization method there are Detection-Based Tracking (DBT) algorithms or Detection-Free Tracking (DFT) algorithms [8]. The first type uses detection algorithms to discover the objects in each of the frames. On the other hand, DFT algorithms require manual selection of the objects in the first frame; afterward, the algorithms identify those objects in the subsequent frames. DBT is much richer in output terms since it can detect new objects in the frames sequence. Another classification of MOT approach is to take into account how the algorithms use the information while processing. Hence, it is possible to observe online or offline tracking [8]; online tracking uses only the actual or past frames to idnetify or re-identify the detectec objects (real-time processing), and offline tracking can use all the sequence of frames, past, present, and future frames (posterior processing).

Nonetheless, something that all the different classifications in the scientific literature [8] have in common is the structure of a MOT algorithm. This structure is illustrated in Fig. 2. Every MOT process starts with a video or a sequence of frames. First, a detection algorithm detects the objects in the different frames; usually, it returns a bounding box per each object detection. The objects detected are related to the detector of the MOT, that is to say, if the MOT aims to track cars, the object detector will identify cars. Second, those detections are analyzed by the tracking algorithm; this phase usually extracts the features to search an affinity with previous frames detected objects (in the case of online processing) and/or with previous or later frames detected objects (in the case of offline processing) in an identification sub-task and associated those identifications over the different frames to comfort the objects paths. This affinity search leads to the association of existing and new objects between frames; in other words, it returns the track of the objects. It is important to clarify that these two MOT processes can be performed separately, first the detection of all frames and then the tracking of all frames (offline processing), or these tasks can be performed repeatedly for each frame (online processing). However, in recent years, research is being carried out on simultaneous detection and tracking for each of the frames [9–12]. We will explain these novel techniques in 3.1.3.

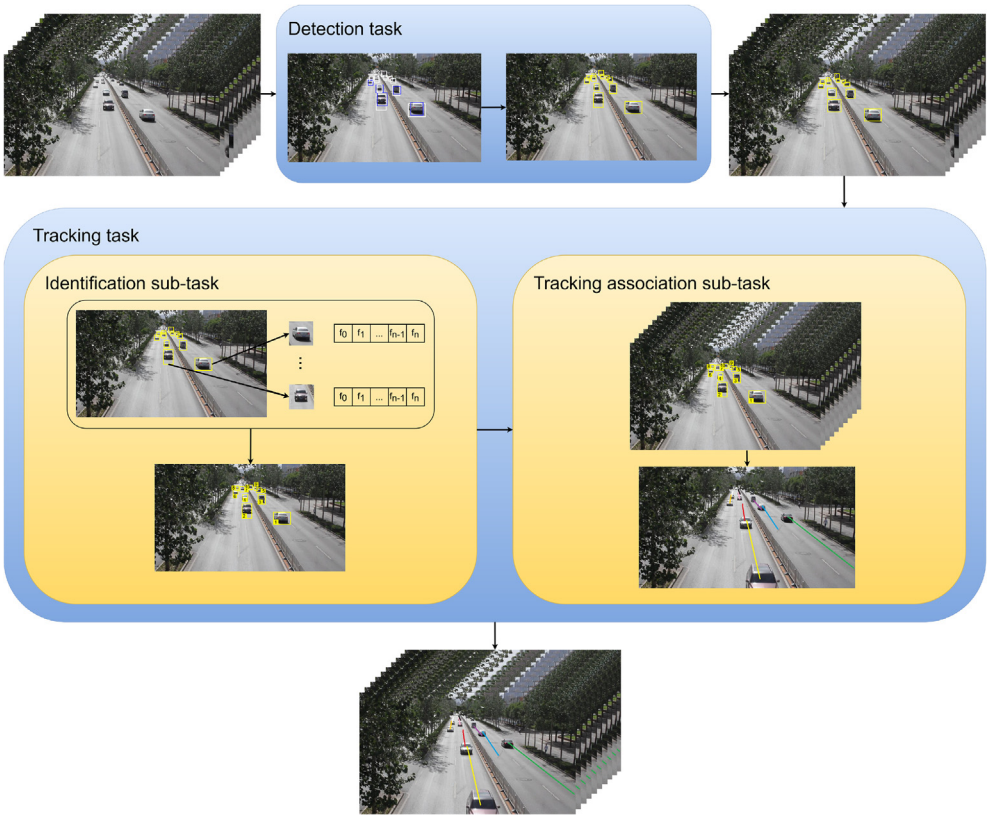

Fig. 2. MOT process' structure.

Now that the main structure of MOT systems is clear let's analyse the different methods and/or techniques used in MOT in urban environments. We will analyse the techniques used for the detection task and afterward the methods used in the tracking part. The resume is available in Fig. 3.

3.1.1. Detection techniques

As we previously explained detection aims to detect the objects of interest in the different frames. DBT systems have to locate all the objects and DFT system only those selected in the first frame. However, both types use the same set of techniques to detect the objects. Our research is focused on urban environments, so the different studies include in this SLR try to identify cars, trucks, motorbikes, vans, cyclists, pedestrians, etc.

The current state of the art techniques for object detection in images is based on Convolutional Neural Networks (CNNs). They are a subtype of artificial neural networks (NNs) and they are applied to different research domains. They are based on convolutional kernels or filters with weights that slide over the input features to provide feature maps. These feature maps allow to identify borders and detect where an object can be. Hence, CNNs offer the final detections of the studied objects.

It is well known that there are several architectures of CNN; however, the most famous one is probably the YOLO algorithm [4]. The performance of this technique is outstanding, and it requires fewer resources than other CNN methods, such us, Faster R-CNN, Fast R-CNN, . . .Especially version number three (YOLOv3) is used in these types of studies for the detection task. This can be seen in [13–18]. Furthermore, Zhang et al. [19] propose to use YOLOv3 along a PointRCNN (Region based Convolutional Neural Network) to detect objects with higher accuracy.

Nonetheless, other kinds of CNNs are commonly used in this type of problem. For instance, R-FCN (Region-based Fully Convolutional Network) is another architecture usually used for object detection [20]. Ooi et al. [21] use this type of CNN to obtain the detections of urban vehicles in their study. Other researchers applied R-FCN along with a non-CNN solution such as background removal [22]; they combined the two techniques to increase the efficiency of the detection task. Also, another powerful CNN is Faster R-CNN; it has been used by Yu et al. [23] and Liu et al. [24]. Faster R-CNN is one of the most powerful architectures for object detection however it usually requires high-efficiency resources. The resources will increase if you use three Faster R-CNN along with Inception-ResNet-v2 as Gunduz and Acarman [25] did or with a two-stage Faster R-CNN [26].

Moreover, we observe that other kind of CNNs are used for this task, for instance, DenseNet [27], ResNeXt101 [28], MobileNetV3 [29], RetinaNet [30], Bi-LSTM (Long Short-Term Memory, this is a Recurrent Neural Netowrk (RNN) not a CNN) [31], and CMNet (Connect-and-Merge convolutional neural network) [32].

On the other hand, another interesting technique used is background removal. We have observed several studies that use this technique to detect objects in a scene. A key example is [33] which uses ViBe background removal for the detection task. Ooi et al. [34] use the IMOT background removal along with PAWCS (Pixel-based Adaptive Word Consensus Segmenter) to polish the object detection. Chandrasekar and Geetha [35] use a fast multi-object tracking method using the Three-Frame Differencing Combined-Background Subtraction (TFDCBS)-coupled-automatic and a fast Histogram-Entropy-Based Thresholding (HEBT) method together with GMPFM-GMPHD filters and a VGG16-LSTM classifier. This last case also uses a CNN as previously explained, however, the background removal technique slightly increases the performance.

Finally, we observe two more used techniques to detect vehicles in a scene, Hidden Markov Models (HMM) [36] and PONO (POsition NOrmalization) [31]. It is important to mention that this last tech-

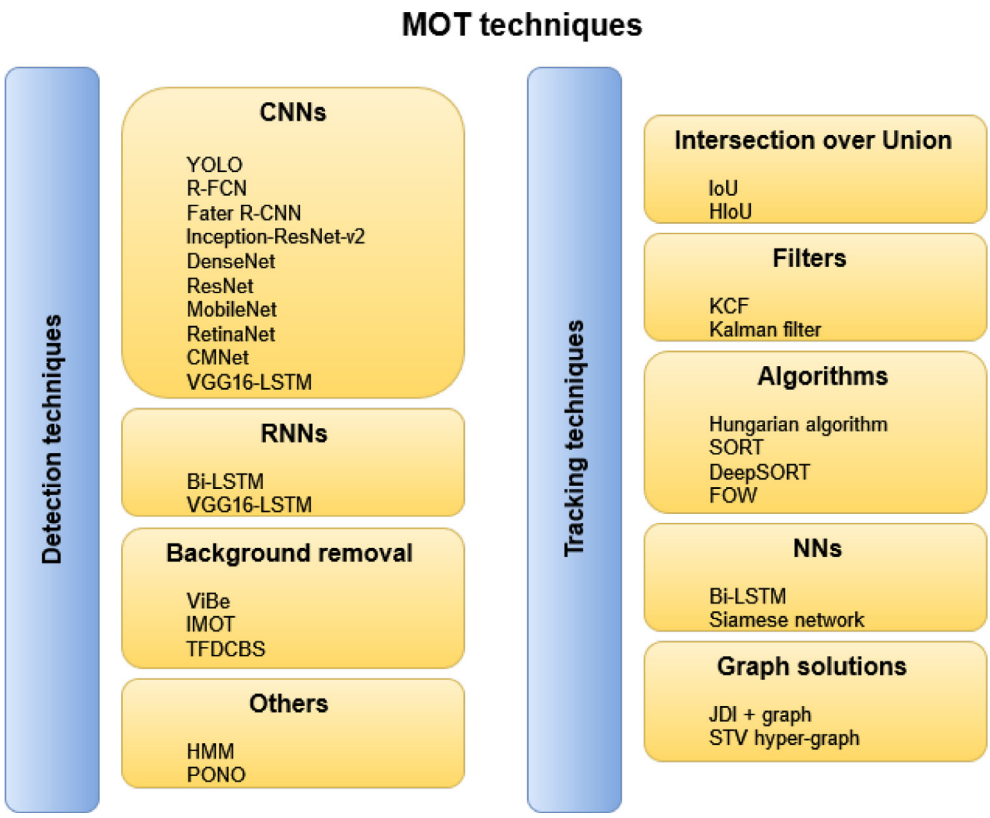

Fig. 3. MOT techniques.

nique is combined with the previously mentioned Bi-LSTM. Therefore, this technique will focus on the recent features of the objects to detect them.

3.1.2. Tracking techniques

As we previously explained tracking techniques aim to related detections in different frames which belong to the same object. Online tracking will only take into account the previous detections, but offline detection will take into account all detections over all the frames. In this subsubsection, we will describe the techniques used in the included papers of this literature review.

The most used method among the included works is Intersection over Union (IoU) tracker. This method can detect when two detections in different frames are related by only taking into account the previous information. It is based on a identification hypothesis and relay on the associations made in the previous iterations. As we can see in these studies [27,19,28,25] is commonly used as a tracking method. It is interesting to point out the study developed by Hua and Anastasiu [13]; they develop the history-based IoU (HIoU). HIoU increases the effectiveness of IoU since it can make a relation with objects that have not been detected due to occlusion. Liang et al. [16] also increase the efficiency of the single IoU method by combining a multi-feature tracking algorithm based on a KCF (Kernel Correlation Filter). This last method can be used to track objects; for instance, the study developed by Yang et al. [37] used a KCF as their tracking algorithm.

Moreover, there is another commonly used method in tracking tasks according to the included articles, is the Kalman filter [38]. As we may know the Kalman filter can estimate the future state according to the previous features. It is used in [36,39,32,17,40]. Furthermore, the Hungarian algorithm is widely used in this type of studies as we can see in [34,22]. It is even used in combination with the Kalman filter [21] to increase metrics. However, this is not the only combination of the Kalman filter with the Hungarian algorithm; Wang and Liu [18] combined both of them along with IoU.

Other tracker algorithms used in the literature are the SORT (Simple, Online, and RealTime) [41] and DeepSORT [42]. However, the SORT algorithm is just the above combination of the Kalman filter and the Hungarian algorithm. Hence, the study developed by Federov et al. [26] also uses this tracker method. On the other hand, DeepSORT adds a pre-trained neural net to obtain objects features that can be used to associate the same objects over different frames as the following studies had demonstrated [24,43,29,30].

Another technique is the Field Of View (FOV). Two included research used this technique as their tracking algorithm [44,45]. This technique delimits an area inside the image and that area is also divided into smaller areas, depending on the enter and exit zone of that area and subareas, the algorithm can track the direction of the detected objects. It is a very basic technique, but it is useful in road intersections and/or when you do not need to know the exact position of the detected elements, but you need to know the enter and exit zone.

Moreover, there are solutions based on NNs like Siamese Network [14] and Bi-LSTM [31] networks. Another solution is to use graph solutions to track objects between the frames, for instance, Tian et al. [46] use the Join Domain Information (JDI) along with a graph-based solution, and Wen et al. [47] use a Space–Time-View (STV) hyper-graph. Other researchers have used the Markov decision process [48], trajectory clustering [49], and even brute force [33].

3.1.3. Simultaneous techniques

Recent studies in the field of MOT focus on performing detection and tracking tasks simultaneously (also known as join techniques) to avoid some of the problems present in this field (see 3.5) and to perform near real-time tracking, thus decreasing the

processing time and making the response faster for users. Some of the most relevant studies and their techniques are explained below.

Zhou et al. [9] have developed a simultaneous MOT process based on point-based tracking. Their process is online and is based on using the centroid of the detections together with a small offset to associate the different detections across the different frames. Furthermore, their detection process is based on the previous frames to improve the results. However, their technique does not take into account when an object reappears in the image as it is assigned a new identifier.

Other researchers, Wang et al. [10], propose that the detection and creation of embeddings should be done simultaneously and shared. In this way, this model can perform these two tasks together to utilize that information with a fast association method that results in the tracking of objects in near real-time. The results obtained by this technique are therefore faster and with similar metrics to the state-of-the-art models.

Wu et al. [11] propose a similar solution to the one already explained in [9]. They make use of tracking offset to improve detection based on the characteristics of the previous frames and detections. They also propose to merge the detection and tracking tasks. Similarly, FairMOT, proposed by Zhang et al. [12], makes use of these strategies in combination with CenterNet and re-ID techniques to obtain the tracks of different objects.

Peng et al. [50] have succeeded in developing a solution that integrates object detection, feature extraction, and data association into a single solution called Chained-Tracker. This solution stands out because it outperforms current solutions in terms of metrics.

The Technical University of Munich, however, proposes a quite different MOT system [51]. They propose a system that has not been trained with tracking data and predicts the next state of the objects based on the detections of the current frame. The solution has proven to be effective in simple tracking environments.

3.2. Which are the devices used for Multi-Object Tracking in traffic environments?

As we have seen in the previous subsection, many of the techniques required high resources to compute and process the data. MOT datasets are commonly composed of video captures which imply high volumes of data that have to be stored and process to train the model. However, we will discuss the different datasets and their characteristics in the next subsection. Nonetheless, these volumes of data are related to the high computing resources needed to resolve MOT problems.

The first important aspect to handle this data is to have an enought amount of RAM that can load the data and process it in real-time while training and configuring the different models. This feature is quite related at the same time with the techniques used for the MOT task. In other words, if the model or technique requires much RAM the task will require much RAM.

During the review process we have observed that most of the studies evaluated mentioned that they had at least 16 GB of RAM [13,36,44,45,49], 32 GB of RAM [31,23,29,48,26], 64 GB of RAM [17], or 128 GB of RAM [47,28]. Devices with 64 or 128 GB of RAM are probably servers' devices; the use of servers is commonly used by computer science researchers to train and evaluate their models and architectures. Another interesting point is that servers usually have more computational resources to deal with computer vision tasks among others. On the other hand, we have noticed that other studies used less RAM, such us, [33] that used 4 GB, [39,46] that used 8 GB, or [35] with 14 GB of RAM.

Furthermore, another important issue to be taken into account when confronting MOT is the frequency of the computer processor. The frequency defines the total number of times that the processor can be activated per second. Hence, as higher the value as higher the frequency. This feature has relevant importance especially when the processor has to make many operations when training the model. As a result, this characteristic can help to reduce the training and testing tasks; an important issue when addressing problems with a huge amount of data.

We have decided to show the different frequencies in several intervals. Hence, it will be easier to analyse the different options. There is only one study [33] included in the interval [0, 2) GHz, 6 studies [36,?,?,?,?,?] in the interval [2, 2.5) GHz, 4 studies [13,28,47,25] belong to the interval [2.5, 3) GHz, 4 studies [15,23,16,26] included in the interval [3, 3.5), 5 studies [35,31,49,39,46] in [3.5, 4), and 2 studies [37,48] included in [4, $+\infty$]. This information is also illustrated in Fig. 4.

Another important component that usually helps to develop and train computer vision models is the use of a GPU. GPU computing significantly increases the velocity of the computing process; to do that, the code has to be prepared for GPU computing. Of course, there are differences between different GPUs in performance and memory. However, for this type of problem is necessary to have GPUs with high memory.

During our review process, we have observed that not all the scientific articles made use of GPU during MOT. From the total 35 papers included only 12 [13,15,28,31,23,24,16,29,30,17,48,26] mentioned that they have made use of a GPU (see Fig. 5).

This analysis of some of the components used in the studies together with the techniques explained in Section 3.1 allows us to establish which of the systems presented in the reviewed articles are intended to be deployed in cloud computing architectures or in edge computing architectures. Thus, the following articles are clearly focused on the edge paradigm [13,36,14,44,33,35,15,45,49,39,37,16,46]; while these other researches focus on cloud services [28,31,23,29,30,17,47,48,26].

3.3. Which are the main datasets for Multi-Object Tracking in traffic environments?

One of the key aspects of every computer science research are the reference datasets in the area for testing and evaluating their proposed models. Hence, MOT algorithms relay their effectiveness not only in the selected model or algorithm but also in the dataset used for training and testing. The datasets are a crucial part of the development of a system; they allow researchers to compare the effectiveness of their proposed models against the effectiveness of previous models with the same data.

Consequently, there are several well-known datasets used for MOT studies. These datasets usually allow performing detection

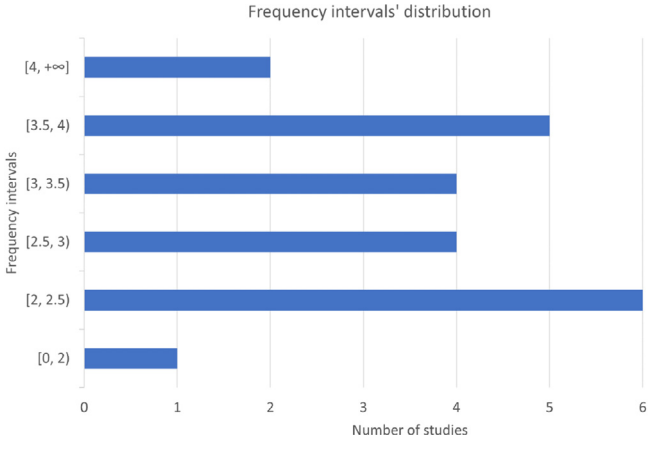

Fig. 4. GHz frecuency intervals' distribution.

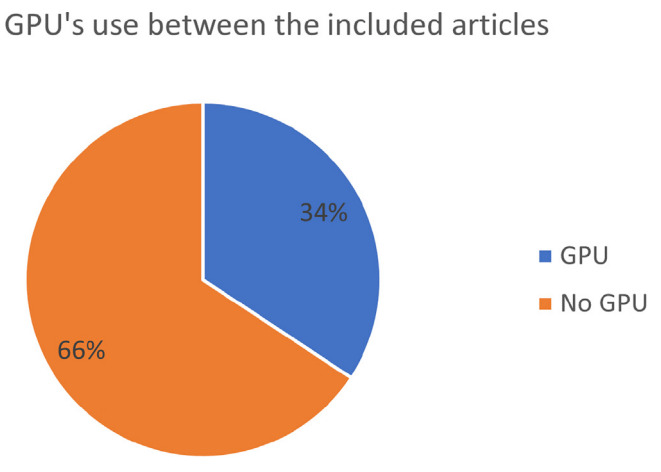

Fig. 5. GPU's use between the included articles.

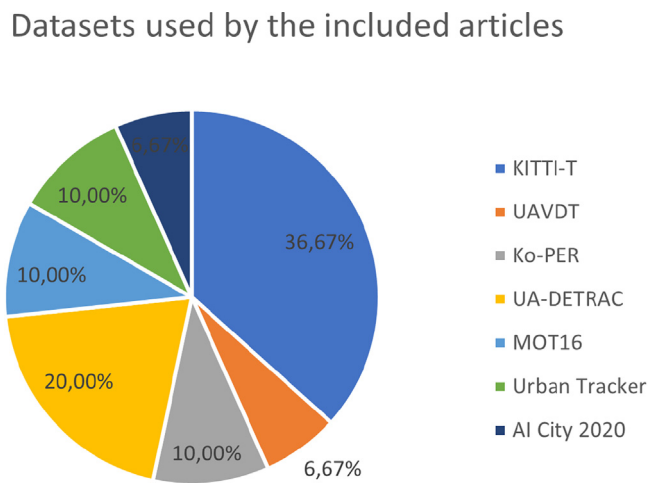

Fig. 6. Datasets used by the included articles.

and tracking activities with several videos recording. Some of the most important datasets are: MOT15 [52], MOT16 [53], MOT20 [54], BU-TIV [55], TUD [56] and PETS2009 [57]. However, only MOT16 and BU-TIV are partially prepared for MOT in traffic environments; they have some sequences of videos in which the datasets are prepper for vehicles detection and tracking. Therefore, these datasets may not be the ideals for performing a MOT in traffic environments.

Nevertheless, the research community has released two MOT datasets oriented to vehicles identification. These datasets are the following: Urban Tracker [58], Ko-PER [59], UAVDT [23], AI City 2020 [60], KITTI-T [61], nuScenes [62], Waymo [63], and UA-DETRAC [64]. Usually, these last two datasets are the most used in studies involving vehicle tracking. We have summarized the most important features of the mentioned datasets in Table 1.

The studies included in the review use the following datasets, KITTI-T, UAVDT, Ko-PER, UA-DETRAC, MOT16, Urban Tracker, nuScenes, and AI City 2020. Fig. 6 shows the use of the different datasets in percentages.

3.4. Which are the evaluation metrics for Multi-Object Tracking in traffic environments?

Currently, the best way to evaluate an Artificial Intelligence (AI) system is to use a metric that illustrates to us how much efficiency and precision the system has. Nevertheless, depending on the type of system we are developing and the type of problem we are planning to resolve, we must pay attention to different evaluation metrics. Consequently, the selection of the correct metric is not a trivial question. In this subsection, we what to know the evaluation metrics commonly used for MOT systems.

During the review process of this article, we had found many papers that do not make use of any evaluation metric [44,30,18,26]; this is not ideal, since every AI system must be evaluated and even compared to other similar solutions to completely check its performance. Another, issue that we had observed is that many researchers use variables, such as, running time or memory use to evaluate the system; these metrics can be used to measure the effectiveness of a model, they only measure how efficient the model is in terms of time and memory. However, we cannot forget these variables, especially when we are implementing a system in a real environment with limited resources.

On the contrary, many research articles make use of evaluation metrics that are not appropriate for MOT problems. These metrics are accuracy or mAP (mean Average Precision) among others. These variables can be useful for the detection sub-task but, they are not for the general MOT task.

Nonetheless, the vast majority of works use other metrics especially designed for MOT systems. They are divided into three main groups explained in detail below:

- **VACE metrics**: they were proposed by Wu and Nevatia [65] and they measure different types of errors in MOT algorithms. The most relevant VACE metrics are the following ones:
  - **FP**: the total number of False Positives in the whole video. FPs are determined by the ground-truth bounding boxes that cannot be associated with any hypothesis.

Table 1
Datasets' summary in traffic environments.

Dataset	Year	Properties				Training set			Testing set		
		Task	Objects	Illumination	Occlusion	Frames	Tracks	Boxes	Frames	Tracks	Boxes
MOT16	2016	Tracking	Pedestrians, vehicles	✓	✓	5.3*k*	467	110*k*	5.9*k*	742	182*k*
BU-TIV	2014	Tracking	Pedestrians, vehicles	-	-	-	-	-	6556	-	-
KITTI-T	2014	Tracking	Vehicles	-	✓	8*k*	-	-	11*k*	-	-
UA-DETRAC	2015	Detection, tracking	Vehicles	✓	✓	84*k*	5.9*k*	578*k*	56*k*	2.3*k*	632*k*
Urban Tracker	2014	Detection, tracking	Vehicles	✓	✓	8.1*k*	-	-	8.1*k*	-	-
Ko-PER	2014	Tracking	Vehicles	-	-	-	-	-	-	-	-
UAVDT	2016	Detection, tracking	Vehicles	-	✓	70*k*	-	-	70*k*	-	-
nuSccenes	2019	Detection, tracking	Pedestrians, vehicles	-	✓	40*k*	-	1.4*M*	40*k*	-	1.4*M*
Waymo	2019	Detection, tracking	Pedestrian, vehicles, signs	-	✓	390*k*	-	12.6*M*	390*k*	-	12.6*M*
AI City 2020	2020	Tracking	Vehicles	✓	✓	190*k*	1.3*k*	-	190*k*	1.3*k*	-

– **FN**: the total number of False Negatives in the whole video. The FNs correspond to those hypotheses that cannot be related to any real bounding box.
– **MT trajectories**: Mostly Tracked trajectories, number of ground-truth trajectories that are correctly tracked in at least 80% of the video's frames.
– **Fragments**: total number of trajectory hypothesis that covers at most 80% of the video's frames. One same trajectory can be covered by more than one fragment.
– **ML trajectories**: Mostly Lost trajectories, number of ground-truth trajectories that are correctly tracked in less than 20% of the video's frames.
– **False trajectories**: predicted trajectories that do not cover the trajectory of a trackable object.
– **ID switches**: number of times that an object is correctly detected but, the ID has been reasigned incorrectly.

- **CLEAR metrics**: they were developed for the Classification of Events, Activities, and Relationships (CLEAR) workshops organized in 2006 and 2007 [66]. These metrics are based on some VACE metrics such as FP, FN, Fragments, and ID switches (IDSW). Nonetheless, to obtain CLEAR metrics we must take into account another metric, IoU (Intersection over Union). IoU helps us to determine when an element and a prediction are related or not. That is to say, if an object o_i and the hypothesis h_i are matched in the frame $t-1$ and on frame t the $IoU(o_i, h_i) >= 0.5$ then o_i and h_i are matched in frame t, even if there is another hypothesis h_j in which $IoU(o_i, h_i) < IoU(o_i, h_j)$. This process continues for every remaining object and hypothesis. Hence, taking that into account and the definitions of the mentioned VACE metrics, the definition of the CLEAR metrics are the followings:
  – **MOTA**: the Multiple Object Tracking Accuracy measures the accuracy in a MOT algorithm. It can be calculated following Eq. 1.

$$MOTA = 1 - \frac{(FN + FP + IDSW)}{GT} \in (-\infty, 1] \tag{1}$$

  where GT represents the number of ground truth boxes.
  – **MOTP**: the Multiple Object Tracking Precision measures the precision in a MOT algorithm. It is calculated by Eq. 2. One important issue about this metric is that it is more focused on the detection's quality than on the tracking task.

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_{t} c_t} \tag{2}$$

  where c_t is the number of matches in frame t, and $d_(t,i)$ is the bounding box that overlaps the hypothesis i.
- **ID Scores**: they were proposed as a complement to CLEAR metrics [67]. They aim to reward those trackers that can follow the same object without ID switches for the longest time. As a consequence, ID scores are obtained differently; they construct a bipartite graph with two different types of vertices. Afterward, considering the connections between the vertices [67] we can obtain True Positive ID (IDTP), True Negative ID (IDTN), False Positive ID (IDFP), and False Negative ID (IDFN). These metrics allow to calculate of the followings ID scores:
  – **Identification precision (IDP)**: is obtained in Eq. 3.

$$IDP = \frac{IDPT}{IDTP + IDFP} \tag{3}$$

  – **Identification recall (IDR)**: is the result of Eq. 4.

$$IDR = \frac{IDPT}{IDTP + IDFN} \tag{4}$$

  – **Identification F1 (IDF1)**: is obtained following Eq. 5.

$$IDF1 = \frac{2IDPT}{2IDPT + IDFP + IDFN} \tag{5}$$

Along with the articles studied for this review, we observe that there are some of these metrics that are more frequently used than others. For instance, CLEAR metrics are the most used ones; they are used in 29 of the 35 studies [13,36,27,14,19,33,68,35,15,28,31,23,39,34,37,24,22,43,16,69,29,46,32,17,21,47,48,25,40]. They are followed by the VACE metric with a total of 20 [13,36,27,14,19,68,35,15,45,28,31,23,39,37,43,69,29,46,25,40] and by the ID Scores metrics with only 2 studies [14,23]. In Fig. 7 we can observe the combination of the metrics used in the different papers.

3.5. What are the main open lines of research or issues in this domain?

Recent advances in the area of computer vision have intensified work in this area and have therefore led to the emergence of new challenges and lines of research. We can see that from the open lines of research mentioned by the included papers.

Hence, one of the most mentioned open lines is to upgrade the tracking technique to handle errors in tracking tasks [36]. Nonetheless, these types of errors not only may be related to the tracking technique but also to the detection techniques and errors that occur while detection is performed [13,33,31]. Moreover, this detection errors may occur by other reasons, for instance, light conditions [30], objects occlusion [44], reflections and shadows [14,15], etc.

Further, occlusion is one of the mains problems that researchers in MOT must deal with. Occlusion can happen when a tracked object is hidden by another tracked one/s or when it is behind something else. Therefore, while one or both situations happened the tracked object is lost. The tracking algorithm must take that into account to handle when the object appears again. This task is called reidentification and it is also one of the open lines of research in several studies [27,31,37,26]. As we can see this task is extremely related to the improvement of the tracking technique mentioned before. To have a good occlusion and reidentification technique is crucial for a good performance of MOT, especially in open spaces.

Some researchers [46] proposed to use deep future, spatiotemporal features, and pose information that may help with the reidentification task. In general terms, the use of deep learning to recognize objects previously seen may help to resolve this problem.

On the other hand, others think that occlusion can be handle with more views [47,44] (Multi Camera Multi Target Tracking, MCMTT). In other words, they aim to use more cameras to decrease occlusion. However, this approach may lead to other problems, and it is harder to resolve in computational terms. It is important to avoid tracking or counting the same objects over the different views. Also, there is a big difference if the views are from the same area or if they are over different spots. So, in this last case, it is also important to consider the reidentification task.

Finally, Zhang et al. [19] propose a 3D multitarget tracking as future research; this propose can be interesting to increase the efficiency and probably to decrease occlusion. It can be performed by forming 3D images from 2D images or by using other kinds of sensors, such as Lidars and/or others.

4. Discussion

From the previous section, we can extract some outcomes that make us understand better how research on MOT is doing. In this section, we will discuss the more relevant aspects concerning the research questions previously answered.

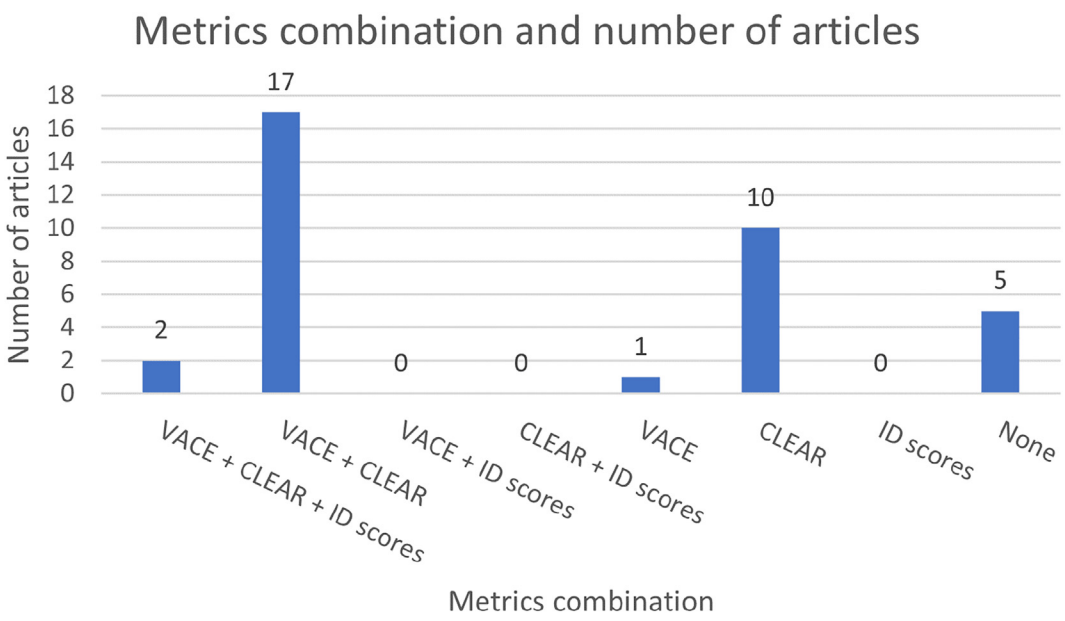

Fig. 7. Metrics combination and number of articles.

When we observe the techniques used in the included papers, we see that the most used method for object detection is the use of deep learning techniques based on CNN. As we mentioned before, CNNs shape the state of the art for object detection in images, therefore, is easy to understand why most of the study use this technique. In general terms, CNNs architectures will detect better relevant objects in the different frames. However, this technique requires more computational cost than other alternatives. In consequence, some researchers decided to use less powerful techniques such as background removal to detect objects. Nevertheless, this technique also has a negative outcome, it is unable to detect non-moving objects. Consequently, the detection and tracking results decrease slightly.

Also, it is important to mention that almost all the tracking techniques can find the associations between objects of the different frames; however, their effectiveness depends on the results of the detection and identification task. That is to say that the accuracy and precision obtained from a MOT algorithm are mostly based on the decision of the detection algorithm and the identification algorithm. Consequently, when developing one of these systems, the selection of the detection techniques is crucial for the whole process.

Another interesting aspect to have in mind, when we are selecting the MOT's algorithms, is the hardware resources that we have to develop the project. It is obvious that if we are going to work with videos and images is important to have enought amount of RAM to infer the models, especially if we are using NNs. Therefore, the selection of the algorithm also depends on the RAM and other factors such as the frequency of the computer processor. A GPU can highly increase the training time and the real-time processing. This feature can be clearly understood when you realize that the articles that used GPU also made use of CNNs.

In the previous section, we have explained the different datasets available and used for MOT in vehicle traffic environments. We have seen that KITTI-T is the most used followed by the UA-DETRAC dataset. These two datasets are probably the most complete ones and offer videos from different perspectives and areas which is useful to train a MOT algorithm under several circumstances. However, the selection of the dataset also depends on where are you going to deploy the system. For instance, the KoPER dataset is specially designed for intersections while the UAVDT is recorded from an Unmanned Aerial Vehicle (UAV). Hence, this dataset will be useful if, for instance, you are planning to control the traffic from a drone or other UAV perspective.

Besides, another interesting research question was which metrics are commonly used when resolving a MOT task in traffic environments. WE have seen that we have three different types of metrics, VACE metrics, CLEAR metrics, and ID Scores. However, when we analyze the use of these metrics in the included articles, we observe that the most used one is the CLEAR metrics. As we mentioned CLEAR metrics are composed of MOTA and MOTP; accuracy and precision are commonly used metrics in classification problems, therefore, it is normal to have and use similar metrics for the tracking problem. Nonetheless, we must not forget that to obtain CLEAR metrics we also need to calculate some VACE metrics. In other words, when we use CLEAR metrics, we also use VACE metrics even if we do not show those metrics in the results section. On the other hand, ID Scores metrics seem not to be too useful for researchers in MOT as the previous metrics are. From our point of view, we consider that the use of CLEAR metrics is the most useful one since it allows you to compare your research with others by using a common metric. Also, CLEAR metrics are easier to understand than ID Scores, since accuracy and precision are terms well-known by the computer science community.

Finally, we must address the open lines and future research in the MOT field. We have seen that MOT in traffic environments state of the art still has some open lines of research. It is crucial to find a good performing algorithm that can handle occlusion and reidentification under different conditions, such as, raining and non-raining, day and night, extreme light conditions, etc. Moreover, some researchers have already proposed some future solutions like multi-view; nevertheless, this approach still has to improve a lot to obtain results that can compete with the single-view solutions. To do that is important to reidentify objects in the different views.

Before finishing this section, we would like to mention that MOT is not only applied in traffic environments for vehicle tracking, traffic condition monitoring at intersections or roads, etc.; it is a technique that can be used in other domains, one of the most studied is MOT on citizens or pedestrians in public spaces; the first MOT studies were applied to this field of study. Although the case studies are different, the techniques used for the detection and

tracking of citizens are practically identical; likewise, the hardware components used to generate this type of model and algorithm are very similar. The metrics used are the same and CLEAR metrics are almost certainly the most widely used in this area of study as well. As for the datasets, there are specific datasets such as [70,71] although some of the datasets mentioned in 1 can also be used for the pedestrian MOT. The problems encountered are very similar, but depending on the environment or scenario where the MOT is developed, it will be easier or more difficult to deal with problems such as occlusion.

Therefore, we can conclude the discussion by saying that MOT in traffic environments still has some aspects to improve and that the work developed by the science community conforms to the state of the art in MOT techniques. Also, we can see that this is a trendy line of research due to the number of research articles developed in recent years and the new datasets that are actualized almost every year to cover new circumstances and to improve the algorithms.

5. Conclusions

A SLR has been carried out in the domain of multi-object tracking in traffic environments. The SLR has been done following the guidelines of Kitchenham and Charters in [6]. The review is focused on the understanding of the MOT problem and the different subareas of this task.

The research covers the techniques used in MOT in traffic environments and explains the main structure of this kind of system. We focus our explanation on the two main tasks of a MOT, detection, and tracking. We explained that the most used technique for detection is based on CNNs architectures and that the effectiveness of the several techniques available for tracking is based on the detection and identifiaction outcome. Moreover, we make a distinction between different types of MOTs.

Afterward, we evaluate the hardware used in the included papers to perform research in the MOT area. These types of research usually imply processing big datasets and training big and complex models. We study and resume the RAM, the frequency of the computer processor, and the use or not of a GPU.

Another issue addressed is the different datasets available and used by the scientific community. We have shown the features of some of the most relevant datasets. This section is useful for new researchers in the area to know the commonly used dataset.

Furthermore, another interesting aspect for new researchers is the metrics used to evaluate the algorithms. We have explained three different metrics and also, we have mentioned the most used one among the researchers. These metrics are really useful to compare the research results with the previous state-of-the-art studies.

To conclude, we also study the direction of this research area for the coming years. The open lines have been mentioned and discussed along with the specific area that has to be developed to put MOT in traffic environments into the next level. This section can be useful to guide new research in this specific area.

To the best of our knowledge, there is no previous SLR in this specific domain and as we had mentioned, we considered that this work can be useful to developers and researchers in MOT in traffic environments. It could be a great initial resource to identify the techniques, the hardware needed, the datasets, the metrics, and the new lines of research.

Funding

This work was supported by the Spanish Agencia Estatal de Investigación. Project Monitoring and tracking systems for the improvement of intelligent mobility and behavior analysis (SiMoMIAC). PID2019-108883RB-C21/ AEI/ 10.13039/501100011033. This work was supported by the Spanish Agencia Estatal de Investigación. Project Monitoring and tracking systems for the improvement of intelligent mobility and behavior analysis (SiMoMIAC). PID2019-108883RB-C22/ AEI/ 10.13039/501100011033. The research of Diego M. Jiménez-Bravo has been co-financed by the European Social Fund and Junta de Castilla y León (Operational Programme 2014-2020 for Castilla y León, EDU/574/2018 BOCYL). The research of André Filipe Sales Mendes has been co-financed by the European Social Fund and Junta de Castilla y León (Operational Programme 2014-2020 for Castilla y León, EDU/556/2019 BOCYL).

CRediT authorship contribution statement

Diego M. Jiménez-Bravo: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft. **Álvaro Lozano Murciego:** Conceptualization, Investigation, Methodology, Software, Supervision, Validation, Writing - review & editing. **André Sales Mendes:** Funding acquisition, Validation, Writing - review & editing. **Héctor Sánchez San Blás:** Formal analysis, Investigation, Validation, Writing - review & editing. **Javier Bajo:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587, https://doi.org/10.1109/CVPR.2014.81, url:https://arxiv.org/abs/1311.2524v5.

[2] R. Girshick, Fast R-CNN, Proceedings of the IEEE International Conference on Computer Vision 2015 (2015) 1440–1448, https://doi.org/10.1109/ICCV.2015.169, url:https://arxiv.org/abs/1504.08083.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6) (2017) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031, url:https://arxiv.org/abs/1506.01497v3.

[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016- (2016), pp. 779–788, https://doi.org/10.1109/CVPR.2016.91, url:https://arxiv.org/abs/1506.02640v5.

[5] K.P. Madore, A.D. Wagner, Cerebrum: the Dana forum on brain science 2019. url:http://www.ncbi.nlm.nih.gov/pubmed/32206165..

[6] S. Keele, Guidelines for performing systematic literature reviews in software engineering, Technical report, Ver. 2.3 EBSE Technical Report. EBSE. url:https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.471..

[7] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation 1 (4) (1989) 541–551, https://doi.org/10.1162/NECO.1989.1.4.541.

[8] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, T.K. Kim, Multiple object tracking: A literature review, Artificial Intelligence 293. doi:10.1016/j.artint.2020.103448. url:https://arxiv.org/abs/1409.7618v4..

[9] X. Zhou, V. Koltun, P. Krähenbühl, Tracking Objects as Points, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12349 LNCS (2020) 474–490. doi:10.1007/978-3-030-58548-8_28. url:https://arxiv.org/abs/2004.01177v2..

[10] Z. Wang, L. Zheng, Y. Liu, Y. Li, S. Wang, Towards Real-Time Multi-Object Tracking, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12356 LNCS (2019) 107–122. doi:10.1007/978-3-030-58621-8_7. url:https://arxiv.org/abs/1909.12605v2..

[11] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, J. Yuan, Track to detect and segment: An online multi-object tracker, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2021, pp. 12347–12356, https://doi.org/10.1109/CVPR46437.2021.01217, url:https://arxiv.org/abs/2103.08808v1.

[12] Y. Zhang, C. Wang, X. Wang, W. Zeng, W. Liu, FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking, International Journal of Computer Vision 129 (11) (2020) 3069–3087. doi:10.1007/s11263-021-01513-4. url:http://arxiv.org/abs/2004.01888..
[13] S. Hua, D.C. Anastasiu, Effective vehicle tracking algorithm for smart traffic networks, in: Proceedings - 13th IEEE International Conference on Service-Oriented System Engineering, SOSE 2019, 10th International Workshop on Joint Cloud Computing, JCC 2019 and 2019 IEEE International Workshop on Cloud Computing in Robotic Systems, CCRS 2019, IEEE; IEEE Comp Soc, IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2019, pp. 67–76. doi:10.1109/SOSE.2019.00019..
[14] Y. Zou, W. Zhang, W. Weng, Z. Meng, Multi-vehicle tracking via real-time detection probes and a markov decision process policy, Sensors (Switzerland) 19 (6). doi:10.3390/s19061309..
[15] J. Wang, S. Simeonova, M. Shahbazi, Orientation- and scale-invariant multi-vehicle detection and tracking from unmanned aerial videos, Remote Sensing 11 (18). doi:10.3390/rs11182155..
[16] H. Liang, H. Song, H. Li, Z. Dai, Vehicle Counting System using Deep Learning and Multi-Object Tracking Methods, Transportation Research Record 2674 (4) (2020) 114–128, https://doi.org/10.1177/0361198120912742.
[17] L. Lou, Q. Zhang, C. Liu, M. Sheng, Y. Zheng, X. Liu, in: Proceedings of 2019 IEEE 8th Data Driven Control and Learning Systems Conference, DDCLS 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 1012–1017. doi:10.1109/DDCLS.2019.8908873. url:https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076438506&doi=10.1109..
[18] K. Wang, M. Liu, YOLOv3-MT: A YOLOv3 using multi-target tracking for vehicle visual detection, Applied Intelligence doi:10.1007/s10489-021-02491-3..
[19] L. Zhang, J. Lai, Z. Zhang, Z. Deng, B. He, Y. He, Multimodal Multiobject Tracking by Fusing Deep Appearance Features and Motion Information, Complexity (2020), https://doi.org/10.1155/2020/8810340.
[20] J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object detection via region-based fully convolutional networks, Advances in Neural Information Processing Systems (2016) 379–387. url:https://arxiv.org/abs/1605.06409v2.
[21] H.L. Ooi, G.A. Bilodeau, N. Saunier, D.A. Beaupré, Multiple object tracking in urban traffic scenes with a multiclass object detector, in: Bebis, G and Boyle, R and Parvin, B and Koracin, D and Turek, M and Ramalingam, S and Xu, K and Lin, S and Alsallakh, B and Yang, J and Cuervo, E and Ventura, J (Ed.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11241 LNCS of Lecture Notes in Computer Science, UNR; DRI; LBNL; NASA Ames; BAE Syst; Intel; Ford; Hewlett Packard; Mitsubishi Elect Res Labs; Toyota; Gen Elect, SPRINGER INTERNATIONAL PUBLISHING AG, GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND, 2018, pp. 727–736. doi:10.1007/978-3-030-03801-4_63..
[22] H.L. Ooi, G.A. Bilodeau, N. Saunier, Tracking in urban traffic scenes from background subtraction and object detection, in: Karray, F and Campilho, A and Yu, A (Ed.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11662 LNCS of Lecture Notes in Computer Science, Assoc Image &s066amp; Machine Intelligence; Univ Waterloo, Fac Engn; Univ Porto, Fac Engn, Dept Elect &s066amp; Comp Engn; Inst Syst &s066amp; Comp Engn Technol &s066amp; Sciand Science; Univ Waterloo, Waterloo AI Inst; Univ Waterloo, Ctr Pattern Anal &s066amp; Machine Intelligence; Inst S, SPRINGER INTERNATIONAL PUBLISHING AG, GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND, 2019, pp. 195–206. doi:10.1007/978-3-030-27202-9_17..
[23] H. Yu, G. Li, W. Zhang, Q. Huang, D. Du, Q. Tian, N. Sebe, The Unmanned Aerial Vehicle Benchmark: Object Detection, Tracking and Baseline, International Journal of Computer Vision 128 (5) (2020) 1141–1159, https://doi.org/10.1007/s11263-019-01266-1.
[24] Z. Liu, W. Zhang, X. Gao, H. Meng, X. Tan, X. Zhu, Z. Xue, X. Ye, H. Zhang, S. Wen, E. Ding, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Vol. 2020-June, IEEE Computer Society, 2020, pp. 2617–2625. doi:10.1109/CVPRW50498.2020.00315. url:https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090154333&doi=10.1109..
[25] G. Gunduz, T. Acarman, Efficient Multi-Object Tracking by Strong Associations on Temporal Window, IEEE Transactions on Intelligent Vehicles 4 (3) (2019) 447–455, https://doi.org/10.1109/TIV.2019.2919473.
[26] A. Fedorov, K. Nikolskaia, S. Ivanov, V. Shepelev, A. Minbaleev, Traffic flow estimation with data from a video surveillance camera, Journal of Big Data 6 (1). doi:10.1186/s40537-019-0234-z..
[27] W. Feng, D. Ji, Y. Wang, S. Chang, H. Ren, W. Gan, Challenges on large scale surveillance video analysis, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Vol. 2018-June of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Comp Soc, IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2018, pp. 69–76. doi:10.1109/CVPRW.2018.00017..
[28] M. Fernández-Sanjurjo, M. Mucientes, V.M. Brea, A Real-Time Processing Stand-Alone Multiple Object Visual Tracking System, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11678 LNCS, Springer International Publishing, 2019, pp. 64–74. doi:10.1007/978-3-030-29888-3_6..
[29] T. Zhang, M. Zhao, Multi-Scale Vehicle Detection and Tracking Method in Highway Scene, in: Proceedings of the 32nd Chinese Control and Decision Conference, CCDC 2020, Chinese Control and Decision Conference, NE Univ; Chinese Assoc Automat Tech Comm Control &s066amp; Decis Cyber Phys Syst; Chinese Assoc Automat; Anhui Univ; IEEE Control Syst Soc; NE Univ, State Key Lab Synthet Automat Proc Ind; Chinese Assoc Automat, Tech Comm Control Theory, IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2020, pp. 2066–2071. doi:10.1109/CCDC49329.2020.9164690..
[30] Z. Wang, B. Bai, Y. Xie, T. Xing, B. Zhong, Q. Zhou, Y. Meng, B. Xu, Z. Song, P. Xu, R. Hu, H. Chai, Robust and fast vehicle turn-counts at intersections via an integrated solution from detection, tracking and trajectory modeling, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Vol. 2020-June, 2020, pp. 2598–2606. doi:10.1109/CVPRW50498.2020.00313..
[31] X. Dong, J. Niu, J. Cui, Z. Fu, Z. Ouyang, Fast Segmentation-Based Object Tracking Model for Autonomous Vehicles, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 12453 LNCS, Springer International Publishing, 2020, pp. 259–273. doi:10.1007/978-3-030-60239-0_18..
[32] M.B. Khalkhali, A. Vahedian, H.S. Yazdi, IEEE Transactions on Intelligent Transportation Systems doi:10.1109/TITS.2021.3050878. url:https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100831591&doi=10.1109..
[33] K. Singh, V. Karar, S. Poddar, Pattern Recognition and Image Analysis 30 (3) (2020) 416–427, https://doi.org/10.1134/S1054661820030268, url:https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091005329&doi=10.1134.
[34] H.L. Ooi, G.A. Bilodeau, N. Saunier, Supervised and unsupervised detections for multiple object tracking in traffic scenes: A comparative study, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 12131 LNCS, Springer International Publishing, 2020, pp. 42–55. doi:10.1007/978-3-030-50347-5_4..
[35] K.S. Chandrasekar, P. Geetha, Multiple objects tracking by a highly decisive three-frame differencing-combined-background subtraction method with GMPFM-GMPHD filters and VGG16-LSTM classifier, Journal of Visual Communication and Image Representation 72. doi:10.1016/j.jvcir.2020.102905..
[36] S. Chattopadhyay, Q. Ge, C. Wei, E. Lobaton, in: 2015 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2015, Institute of Electrical and Electronics Engineers Inc., 2016, pp. 805–809. doi:10.1109/GlobalSIP.2015.7418308. url:https://www.scopus.com/inward/record.uri?eid=2-s2.0-84964794695&doi=10.1109..
[37] H. Yang, S. Gao, X. Wu, Y. Zhang, Online multi-object tracking using KCF-based single-object tracker with occlusion analysis, Multimedia Systems 26 (6) (2020) 655–669, https://doi.org/10.1007/s00530-020-00675-4.
[38] R.E. Kalman, A new approach to linear filtering and prediction problems, Journal of Fluids Engineering, Transactions of the ASME 82 (1) (1960) 35–45, https://doi.org/10.1115/1.3662552.
[39] L. Jun, X. Wei-xin, L. Liang-qun, Online Visual Multiple Target Tracking by Intuitionistic Fuzzy Data Association, International Journal of Fuzzy Systems 19 (2) (2017) 355–366, https://doi.org/10.1007/s40815-016-0172-2.
[40] N.M. Al-Shakarji, F. Bunyak, G. Seetharaman, K. Palaniappan, Multi-object Tracking Cascade with Multi-Step Data Association and Occlusion Handling, in: Proceedings of AVSS 2018–2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance, IEEE; IEEE Signal Proc Soc, IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2019, pp. 423–428. doi:10.1109/AVSS.2018.8639321..
[41] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, Proceedings - International Conference on Image Processing, ICIP 2016-Augus (2016) 3464–3468. doi:10.1109/ICIP.2016.7533003. url:https://arxiv.org/abs/1602.00763v2..
[42] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, Proceedings - International Conference on Image Processing, ICIP (2017- (2018)) 3645–3649, https://doi.org/10.1109/ICIP.2017.8296962, url:https://arxiv.org/abs/1703.07402v1.
[43] X. Hou, Y. Wang, L.P. Chau, Vehicle tracking using deep SORT with low confidence track filtering, in: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2019, IEEE, IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2019. doi:10.1109/AVSS.2019.8909903..
[44] M. Delavarian, O. Maarouzi, in: Proceedings - 3rd Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2017, Vol. 2017-Decem, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 131–135. doi:10.1109/ICSPIS.2017.8311603. url:https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050820485&doi=10.1109..
[45] M. Delavarian, O. Marouzi, H. Hassanpour, A multilayer motion direction based model for tracking vehicles at intersections, International Journal of Engineering, Transactions A: Basics 33 (10) (2020) 1939–1950, https://doi.org/10.5829/IJE.2020.33.10A.12.
[46] W. Tian, M. Lauer, L. Chen, Online Multi-Object Tracking Using Joint Domain Information in Traffic Scenarios, IEEE Transactions on Intelligent Transportation Systems 21 (1) (2020) 374–384, https://doi.org/10.1109/TITS.2019.2892413.
[47] L. Wen, Z. Lei, M.C. Chang, H. Qi, S. Lyu, Multi-Camera Multi-Target Tracking with Space-Time-View Hyper-graph, International Journal of Computer Vision 122 (2) (2017) 313–333, https://doi.org/10.1007/s11263-016-0943-0.
[48] J.E. Espinosa, S.A. Velastin, J.W. Branch, Detection and Tracking of Motorcycles in Congested Urban Environments Using Deep Learning and Markov Decision Processes, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11524 LNCS, Springer International Publishing, 2019, pp. 139–148. doi:10.1007/978-3-030-21077-9_13..
[49] A.A. Sekh, D.P. Dogra, S. Kar, P.P. Roy, Video trajectory analysis using unsupervised clustering and multi-criteria ranking, Soft Computing 24 (21) (2020) 16643–16654, https://doi.org/10.1007/s00500-020-04967-9.

[50] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Fu, Chained-Tracker: Chaining Paired Attentive Regression Results for End-to-End Joint Multiple-Object Detection and Tracking, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12349 LNCS (2020) 145–161. doi:10.1007/978-3-030-58548-8_9. url:https://arxiv.org/abs/2007.14557v1.
[51] P. Bergmann, T. Meinhardt, L. Leal-Taixe, Tracking without bells and whistles, Proceedings of the IEEE International Conference on Computer Vision 2019-October (2019) 941–951. doi:10.1109/ICCV.2019.00103. url:http://arxiv.org/abs/1903.05625 https://doi.org/10.1109/ICCV.2019.00103.
[52] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, K. Schindler, MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. url:http://arxiv.org/abs/1504.01942..
[53] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, K. Schindler, MOT16: A Benchmark for Multi-Object Tracking. url:http://arxiv.org/abs/1603.00831..
[54] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, L. Leal-Taixé, MOT20: A benchmark for multi object tracking in crowded scenes. url:https://arxiv.org/abs/2003.09003v1..
[55] Z. Wu, N. Fuller, D. Theriault, M. Betke, A thermal infrared video benchmark for visual analysis, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2014) 201–208, https://doi.org/10.1109/CVPRW.2014.39.
[56] C. Morris, N.M. Kriege, F. Bause, K. Kersting, P. Mutzel, M. Neumann, http://arxiv.org/abs/2007.08663TUDataset: A collection of benchmark datasets for learning with graphs. url:http://arxiv.org/abs/2007.08663..
[57] J. Ferryman, A. Shahrokni, PETS2009: Dataset and challenge, Proceedings of the 12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS-Winter 2009 doi:10.1109/PETS-WINTER.2009.5399556..
[58] J.P. Jodoin, G.A. Bilodeau, N. Saunier, Urban Tracker: Multiple object tracking in urban mixed traffic, 2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014 (2014) 885–892 doi:10.1109/WACV.2014.6836010..
[59] E. Strigel, D. Meissner, F. Seeliger, B. Wilking, K. Dietmayer, The Ko-PER intersection laserscanner and video dataset, 2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014 (2014) 1900–1901 doi:10.1109/ITSC.2014.6957976..
[60] M. Naphade, S. Wang, D.C. Anastasiu, Z. Tang, M.C. Chang, X. Yang, L. Zheng, A. Sharma, R. Chellappa, P. Chakraborty, The 4th AI city challenge, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2020- (2020)) 2665–2674, https://doi.org/10.1109/CVPRW50498.2020.00321, url:https://arxiv.org/abs/2004.14619v1.
[61] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The KITTI dataset, International Journal of Robotics Research 32 (11) (2013) 1231–1237, https://doi.org/10.1177/0278364913491297, url:https://journals.sagepub.com/doi/full/10.1177/0278364913491297.
[62] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuScenes: A multimodal dataset for autonomous driving, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2019) 11618–11628 doi:10.1109/CVPR42600.2020.01164. url:https://arxiv.org/abs/1903.11027v5.
[63] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, D. Anguelov, Scalability in Perception for Autonomous Driving: Waymo Open Dataset, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2019) 2443–2451, https://doi.org/10.1109/CVPR42600.2020.00252, url:https://arxiv.org/abs/1912.04838v7.
[64] S. Lyu, M.C. Chang, D. Du, L. Wen, H. Qi, Y. Li, Y. Wei, L. Ke, T. Hu, M. Del Coco, P. Carcagni, D. Anisimov, E. Bochinski, F. Galasso, F. Bunyak, G. Han, H. Ye, H. Wang, K. Palaniappan, K. Ozcan, L. Wang, L. Wang, M. Lauer, N. Watcharapinchai, N. Song, N.M. Al-Shakarji, S. Wang, S. Amin, S. Rujikietgumjorn, T. Khanova, T. Sikora, T. Kutschbach, V. Eiselein, W. Tian, X. Xue, X. Yu, Y. Lu, Y. Zheng, Y. Huang, Y. Zhang, UA-DETRAC 2017: Report of AVSS2017 & IWT4S Challenge on Advanced Traffic Monitoring, 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017 doi:10.1109/AVSS.2017.8078560..
[65] B. Wu, R. Nevatia, Tracking of multiple, partially occluded humans based on static body part detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1, 2006, pp. 951–958, https://doi.org/10.1109/CVPR.2006.312.
[66] R. Stiefelhagen, J.S. Garofolo, Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006 Revised Selected Papers, Vol. 4122 LNCS, Springer, Southampton, 2007. url:https://www.researchgate.net/publication/239577438_Multimodal_Technologies_for_Perception_of_Humans_First_International_Evaluation_Workshop_on_Classification_of_Events_Activities_and_Relationships_CLEAR_2006_Southampton_UK_April_6-7_2006_Revised_Selec..
[67] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9914 LNCS (2016) 17–35. doi:10.1007/978-3-319-48881-3_2. url:https://link.springer.com/chapter/10.1007/978-3-319-48881-3_2..
[68] I. del Pino, V. Vaquero, B. Masini, J. Solà, F. Moreno-Noguer, A. Sanfeliu, J. Andrade-Cetto, Low resolution lidar-based multi-object tracking for driving applications, Advances in Intelligent Systems and Computing, Vol. 693, Springer International Publishing, 2018, pp. 287–298, https://doi.org/10.1007/978-3-319-70833-1_24.
[69] L. Wen, D. Du, Z. Cai, Z. Lei, M.C. Chang, H. Qi, J. Lim, M.H. Yang, S. Lyu, UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking, Computer Vision and Image Understanding 193. doi:10.1016/j.cviu.2020.102907..
[70] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2012) 743–761, https://doi.org/10.1109/TPAMI.2011.155.
[71] J.F.P. Kooij, N. Schneider, F. Flohr, D.M. Gavrila, Context-based pedestrian path prediction, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8694 LNCS (PART 6) (2014) 618–633. doi:10.1007/978-3-319-10599-4_40..

Diego M. Jiménez-Bravo studied a degree in Computer Engineering Management and Information Systems at the University of the Basque Country (UPV/EHU) (2016). Subsequently, he obtained a Master's degree in Intelligent Systems from the University of Salamanca (USAL) (2017). He concluded his formative stage by obtaining a Ph.D. in Computer Engineering from the USAL (2020). He is part of the research group ESALab, of the same entity, where he carries out research work under a contract funded by the Junta de Castilla y León and the European Social Fund. His main research interests focus on the field of artificial intelligence, IoT (internet of things), smart-homes, energy optimisation and social networks, among others. He also collaborates in several research projects within the research group.

Á**lvaro Lozano Murciego** received the master's degree in intelligent systems and the Ph.D. degree in computer engineering from the University of Salamanca, in 2015 and 2019, respectively. He is currently working on the Expert Systems and Applications Laboratory Research Group, Computer Science Department, University of Salamanca, as an Assistant Professor. Throughout his training, he has followed a well-defined line of research, focused on machine learning, route optimization, IoT sensors, and edge computing.

André Sales Mendes is a Ph.D. student in computer engineering. He has completed a computer engineering degree and a master's degree in intelligent systems at the University of Salamanca. He is currently focusing his doctoral studies on artificial intelligence and expert robots. He is a member of the Expert Systems And Applications Laboratory research group at the University of Salamanca. In addition, he has several publications in JCR indexed journals.

Héctor Sánchez San Blas is a Ph.D. Student in computer engineering. He has completed a Computer Engineer degree and a Master's degree in Intelligent Systems at the University of Salamanca. During the master's degree, he was a collaborating researcher at the Expert Systems and Applications Lab (ESALab) at the same university, collaborating with research projects related to the Internet of Things, Virtual Reality applications, and machine learning. Currently, he researches IoT and neural networks development and research projects focused on artificial vision and Smart Cities.

Javier Bajo is a Professor at the ETS Ingenieros Informáticos of the Universidad Politécnica de Madrid. He is currently Director of the Research Center in Artificial Intelligence of the UPM. Previously he was a professor at the Pontifical University of Salamanca (2003 to 2012) and Director of the Data Processing Center of that University (2010–2012).He holds a Ph.D. in Computer Science from the University of Salamanca (with honors) (2007) and a Master in Electronic Commerce from the same University (2006). His lines of research focus on multi-agent systems, social computing and knowledge representation and reasoning. He has participated in more than 50 research projects (European, National or Regional) and research contracts, being principal investigator in 11 of them.