applied sciences

MDPI

Review

Pedestrian and Cyclist Detection and Intent Estimation for Autonomous Vehicles: A Survey

Sarfraz Ahmed [1,*], M. Nazmul Huda [1], Sujan Rajbhandari [1], Chitta Saha [1], Mark Elshaw [1] and Stratis Kanarachos [2]

[1] School of Computing, Electronics and Mathematics, Coventry University, Coventry CV1 5FB, UK; ab9467@coventry.ac.uk (M.N.H.); ac1378@coventry.ac.uk (S.R.); ab3135@coventry.ac.uk (C.S.); ab0487@coventry.ac.uk (M.E.)

[2] School of Mechanical, Aerospace and Automotive Engineering, Coventry University, Coventry CV1 5FB, UK; ab8522@coventry.ac.uk

* Correspondence: ahmed157@uni.coventry.ac.uk

Received: 12 April 2019; Accepted: 3 June 2019; Published: 6 June 2019

Abstract: As autonomous vehicles become more common on the roads, their advancement draws on safety concerns for vulnerable road users, such as pedestrians and cyclists. This paper presents a review of recent developments in pedestrian and cyclist detection and intent estimation to increase the safety of autonomous vehicles, for both the driver and other road users. Understanding the intentions of the pedestrian/cyclist enables the self-driving vehicle to take actions to avoid incidents. To make this possible, development of methods/techniques, such as deep learning (DL), for the autonomous vehicle will be explored. For example, the development of pedestrian detection has been significantly advanced using DL approaches, such as; Fast Region-Convolutional Neural Network (R-CNN) , Faster R-CNN and Single Shot Detector (SSD). Although DL has been around for several decades, the hardware to realise the techniques have only recently become viable. Using these DL methods for pedestrian and cyclist detection and applying it for the tracking, motion modelling and pose estimation can allow for a successful and accurate method of intent estimation for the vulnerable road users. Although there has been a growth in research surrounding the study of pedestrian detection using vision-based approaches, further attention should include focus on cyclist detection. To further improve safety for these vulnerable road users (VRUs), approaches such as sensor fusion and intent estimation should be investigated.

Keywords: pedestrian detection; cyclist detection; deep learning; CNN; Fast R-CNN; Faster R-CNN; pose estimation; motion modelling; tracking; intent estimation

1. Introduction

The rise in the development of autonomous vehicles underpins essential safety concerns particularly for vulnerable road users (VRUs) such as pedestrians and cyclists. Concerns have been mounting specifically surrounding whether the autonomous vehicle is able to take them into consideration while operating on public roads. Therefore, it is critical that the autonomous vehicle can detect, classify and predict the intention of the VRUs in real time, and required action is taken not to compromise the safety of other road users. To achieve this, deep learning (DL) techniques have recently been employed for detection and pose estimation to predict the intention of pedestrians and cyclists. For example, Convolutional Neural Networks (CNNs), a type of DL technique, have been highly successful in the field of object detection, particularly, pedestrian detection [1–5]. Recent advances of such DL techniques have outperformed previous methods of computer vision problems (see [6–10] reduced number of refs). Some DL techniques used for pedestrian detection have achieved miss rates of less than 10% [11]. Although the miss rate is significantly low, they are yet to reach human levels

of detection, and therefore significant research is still necessary. Until detection levels are improved, autonomous vehicles remain a danger to VRUs.

According to the World Health Organisation (WHO), nearly half of road traffic fatalities are experienced by pedestrians and cyclists than any other road users as they do not have any special means of protection (i.e., helmets, clothing, etc.) [12]. To be able to predict the intention of a pedestrian using identification and pose estimation techniques would provide a higher level of safety for all road users. In 2013, WHO reported that it is expected that traffic accidents will be the fifth leading cause of death by 2030, rising from the current eighth position [13,14]. In 2013, VRUs make up more than a quarter of victims of traffic accidents. Of the deaths recorded due to traffic accidents, 42% were pedestrians and 16% were cyclists, with 69% of these fatal accidents occurring in urban locations. In 2017, of all fatalities due to road traffic accidents, 21% were pedestrians, and 8% were cyclists [15]. In the UK, pedestrians and cyclists accounted for 26% and 6% of road traffic fatalities, respectively [16] in 2017. Most accidents occurred in rural roads (55%) and Urban areas (37%). It is also worth noting that half of the accidents involving pedestrians occur at night [17,18].

Autonomous vehicles aim to make the roads safer for the VRUs through accurate detection. Although detection systems have become more accurate, they have yet to reach human levels. To improve the accuracy of detection systems, the challenges that need to be overcome include occlusion, crowding, weather and lighting conditions. The flowchart in Figure 1 represents the tasks required by the autonomous vehicles to safely detect and estimate the future actions of VRUs. This process allows the vehicle to safely navigate with respect to the VRU. The interaction between the autonomous vehicle and its surroundings is achieved via sensors which collect information primarily to detect and track objects. The sensor input in Figure 1 is affected by external sources, which can reduce efficiency. Typically, the sensing method relies on a vision-based approach such as visible band cameras (operating at the spectrum of 400 nm to 700 nm) [1–4]. Sensing based on the visible light spectrum is susceptible to ambient light, shadowing and weather conditions. During low-light conditions due to the time of the day, weather, shadowing, etc. can reduce the accuracy of the sensors. A common approach to overcome this problem is to create multiple sensor systems using sensor fusion (e.g., combining visible and infrared band camera) to increase the robustness and accuracy [3,4]. The thermal sensor detects the thermal radiation from an object, which allows the detection and tracking of pedestrians and cyclists in low-light conditions.

The accuracy of the classification, detection and pose estimation are based on the quality of the sensor data information. The focus of this paper is to provide an overview of the current pedestrian and cyclist detection and intent estimation techniques and compare the existing techniques. Building upon the vast existing literature in the field of computer vision and object detection, pedestrian and cyclist detection will be explored and discussed. The detection stage allows for identification and location of such objects in images and video frames [19], therefore making it a vital part of autonomous vehicles [20–23]. Detection results are then used for tracking and pose estimation of the pedestrians/cyclists. As DL techniques for VRU detection and intent estimation will be the primary focus, this will not encompass tracking techniques.

The purpose of this survey is to provide comprehensive review of the recent studies undertaken in both pedestrian and cyclist detection and pose estimation based on state-of-the-art sensor fusion and DL techniques. There is limited work focused on cyclists detection compared to pedestrian detection. There is also limited work on using multispectral data for VRU detection. Using sensor fusion techniques with DL can lead to improved results based on previous state-of-the-art methods. Therefore, it is critical to find an optimal fusion technique to improve the detection accuracy of the system. Once detected, pose estimation techniques can be applied to the VRUs.

The organisation of the paper is as follows: Section 2 will highlight the challenges and importance of detection and intent estimation for autonomous vehicles. Sections 3 and 4 will provide a brief history of object detection techniques and its typical detection pipeline. Section 5 will explore the state-of-art techniques based on DL currently used in pedestrian and cyclist detection. Section 6 discusses the

architectures of the DL-based detectors for pedestrians and cyclists. Section 7 outlines the datasets used for pedestrian and cyclist detection. Section 8 will discuss DL-based sensor fusion approaches for an improving detection. Section 9 introduces the latest DL approaches that are used for pose estimation and intent estimation. Concluding remarks and future works will be presented in Section 10.

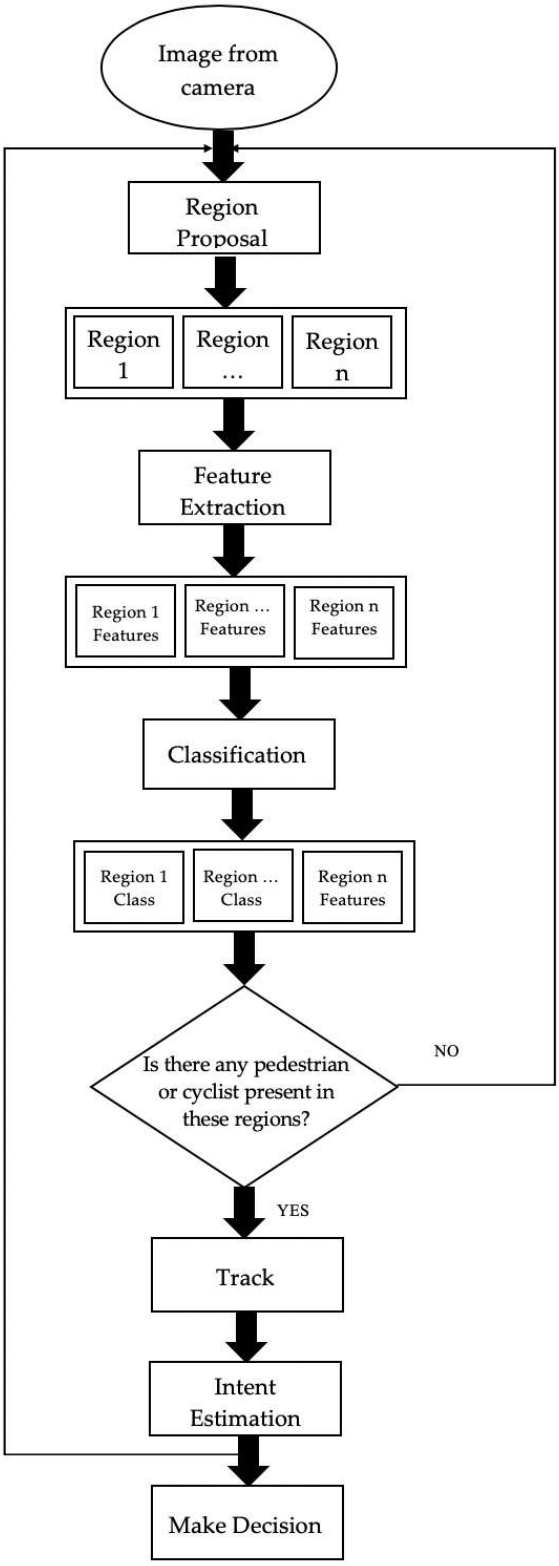


Figure 1. A flowchart of detection and intent estimation system.

2. Challenges of Detection and Intent Estimation

Advanced driver assistance system (ADAS) technology, such as cruise control, emergency braking and lane departure system have brought a certain level of safety for vehicles and other road users. Automatic speed control (cruise control) was developed in the early 1990s, based on electronic cruise control technology that was introduced in the late 1960s. It was not widely implemented until the 1980s [24]. From the cruise control technology, adaptive cruise control was developed. It uses sensors to detect vehicles in front to adjust speed to maintain a distance between the vehicles. These sensors have also been used for emergency braking if an object is detected within a given range. Lane departure systems are used for warning the driver of potentially unintended lane changes. Initially designed for semi-truck drivers, they were adopted by consumer vehicles in 2001 as part of the lane keeping support system by Nissan [24]. By monitoring the use of the indicators, the system detects if a lane departure is intentional. If the vehicle begins to change lanes without the use of indicators, the system warns the driver. The systems discussed above are dependent on driver intervention and focus on a single aspect of dangers on the road. However, the technology discussed above cannot provide a sufficient level of safety for a fully autonomous vehicle. So, further research is required to increase detection accuracy for autonomous vehicles.

For a fully automated vehicle, the detection of dangers associated with pedestrian/cyclist detection should be a continuous operation, as represented in Figure 1. This cannot be achieved by a driver driven vehicle as the driver cannot maintain a continuous level of awareness to their surroundings. Even with the considerable progress on autonomous vehicles, further development is pivotal for pedestrian and cyclist detection to address safety concerns. Therefore, this continues to be an area that is being investigated and explored, as in [25–27] reduced ref grouping.

3. Detection Techniques: A Brief History

Detection techniques, especially for pedestrians, has been widely researched with several techniques. The first instance of object detection is known as the region of interest (ROI) [28]. Once the potential location of the desired object (i.e., pedestrian or cyclist) is identified in an image, feature extraction takes place. These features can include edges, shapes, curvature, etc. These features are sent to a classifier for classification [28] (see Figure 1).

The Background Subtraction (BS) approach was the first technique applied for detecting a moving object. In this approach, the moving objects are identified by comparing the current frame with the reference frame, known as the background image [23,29]. This method is simple to implement but is susceptible to environmental conditions such as light intensities (i.e., time of day, shadowing) and dynamic backgrounds [30]. To improve the detection and tracking, a number of advanced techniques such as the sliding window, objectiveness and selective search were developed [31].

Algorithms for feature extraction and classification for object detection can be either hand-crafted or DL-based methods. Hand-crafted methods for feature extraction are based on models that were manually designed on low-level features to propose ROIs [19]. These models were based on techniques such as BS, the histogram of oriented gradients (HOG) features [32,33] or local binary pattern (LBP) [34]. Hand-crafted methods can be limited and not very robust as complex features can be difficult to hand-craft. DL techniques allow the network to determine features. This can provide a higher level of abstraction.

Then classifiers, such as a Support Vector Machine (SVM) [19,35–38], a decision tree [19,36–38] or a deep network [39,40] are used to classify the object (e.g., pedestrian, cyclist) in the image or video sequence. Deep networks have shown promising results in pedestrian detection, outperforming some traditional methods of pedestrian detection. DL-based techniques will be discussed in later sections.

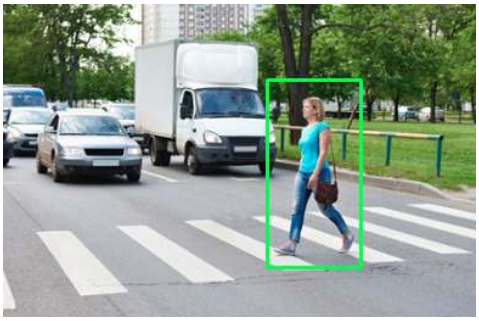
Some of the more commonly used hand-crafted techniques for pedestrian and cyclist detection are discussed below. Haar-like features detect the changes in intensities in the horizontal, vertical and diagonal directions to detect the object [23,41]. Viola and Jones (VJ) implemented the Haar-like features detection approach, while also taking into account the intensity information from the video

frame [30,42]. Introduced in 2003, it used the sudden changes in pixel intensities to detect the shape of an object [42–44]. The VJ detector was one of the earlier techniques designed for pedestrian detection [42]. It used box-shaped filters for feature extraction which is then fed into a classifier based on adaptive boosting known as AdaBoost [45]. Dalal and Triggs presented the HOG (detector which uses a linear SVM for classification [25,32,43,44]). The HOG detector finds an object's shape and appearance based on the intensities of the local gradients or the orientation of the edge [23,32]. The HOG detector became a building block for the Deformable Part Model (DPM) detector in later works [25,35,44,46,47]. DPM was used to weaken the effects of deformation of non-rigid objects [48]. DPM is a popular method for object detection and works well with varying and occluded appearances [48,49]. Based on the DPM, many other object detection methods have been proposed [50]. DPM was implemented in [25] to simultaneously detect and classify both pedestrians and cyclists using an innovative detection approach with a deep network for classification and localisation. The detection method, upper body-multiple potential regions (UB-MPR), focused on the UB of the pedestrian/cyclist for object candidate abstraction as the UB of these road users are normally similar and visible. The potential object regions were extracted using multiple potential regions (MPR) for the UB of the candidate. These potential objects were then sent to a Fast Region-Convolutional Neural Network (R-CNN) [51] for classification. A Fast R-CNN is a DL approach which will be discussed in later sections. A similar approach for using DPM was found in [52]. Methods for detecting pedestrians can be employed for cyclist detection as in [53,54]. LBP uses a neighbourhood of each pixel to extract features [23,34]. This method is very robust compared with the methods above and therefore has become very popular.

4. Typical Detection Pipeline

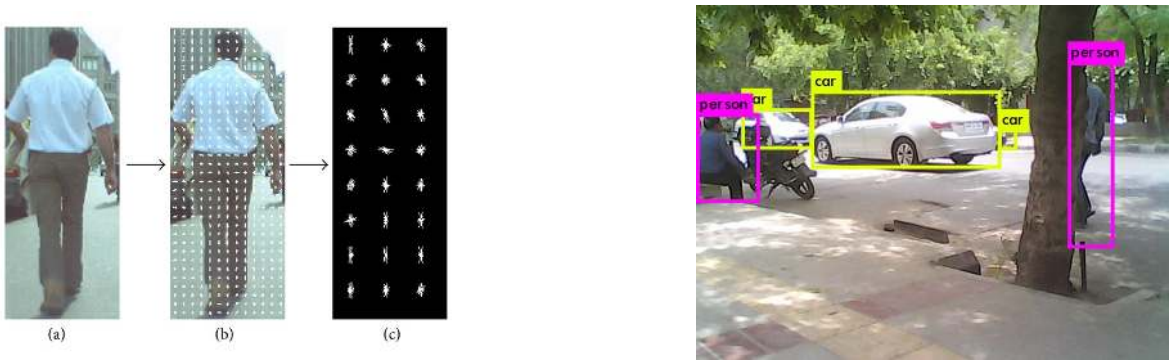
Pedestrian and cyclist detection algorithms mostly follow a basic pipeline or structure (as shown in Figure 2): (a) information collected by the sensor system (b) region of proposals, (c) feature extraction, and (d) classification [23]. These pipelines are described in detail in the following section. The detection pipeline is the first aspects of the overall detection and intent estimation system as described in Figure 1.

(**a**)



(**b**)

Figure 2. *Cont.*



(**c**) (**d**)

Figure 2. Basic Detection System Structure: (**a**) the sensor system captures data in the form of an image (**b**) region proposal techniques are applied, (**c**) features are extracted from the proposed regions and (**d**) finally the features are fed into a classifier. (**a**) Sensor System [55]. (**b**) Region Proposal [56]. (**c**) Feature Extraction [57]. (**d**) Classification [58].

4.1. Regions of Interest

ROIs, also known as region proposal, is regarded as the first and most vital step in a detection system [59]. Some image processing techniques are applied at this stage for ease of finding ROIs [23]. The region proposals have been typically visible-based approaches, such as monocular or stereo cameras. The features such as edges, lines, patterns are then extracted and processed through to the classifier to determine the class of the object (e.g., whether the object is human or not). ROIs are proposed in an image to detect potential pedestrians/cyclists in the scene. Approaches that can be used for finding ROIs include sliding window, selective search [60] and locally decorrelated channel features (LDCF) [44].

In the sliding window approach, a window is scanned both vertically and horizontally to extract candidate regions. These regions may be different scales as the pedestrians can be of varying sizes. No positive regions are discarded as all the regions are fed into CNN. This will provide high accuracy but with a higher level of computational complexity due to the large number false positives [44]. Selective search uses a coarse filter for detecting class-independent regions [60]. This has been successfully used with CNNs for feature extraction and classification [51]. The approach reduces the number of regions proposed, reducing the computational costs. LDCF can detect pedestrians with high accuracy [61]. To further improve, this approach is coupled with a neural network [44], where a large number of regions are produced, each with a confidence value. The confidence value refers to the likelihood that a pedestrian is contained in the frame. This provides for a trade-off of accuracy and efficiency of the detector.

The sliding window approach is the simplest technique and is adaptable for use with various aspect ratios and scales [44]. However, more complex algorithms can lead to lower the number of ROIs, reducing the number of false positives. This also reduces the computational costs of the overall detection system.

4.2. Feature Extraction

Feature extraction of the ROIs is processed. Some of the major and well-known feature extraction techniques were discussed in Section 3. Depending on the application, different techniques can be applied [44]. For example, to identify visible characteristics the VJ descriptor, HOG descriptor and DL approaches can be applied. The VJ descriptor uses intensity contrasts for feature extraction while the HOG descriptor uses pooled gradients. DL techniques can be used when certain features cannot be hand-crafted. For each input region, a vector of real-valued or binary values are produced. The output vector represents the visible characteristics of the proposed regions.

4.3. Classification

The output vector produced from the feature extraction stage is fed into a classifier to determine if a pedestrian or any other object exists in the proposed regions in the form of a binary label. Classifiers that have been used in the previous studies with feature extractors include AdaBoost [42] and SVM [32]. However, with the advancement of DL, more often, CNN-based approaches, are being implemented for classification. These CNN-based approaches will be discussed in the next section.

5. Deep Learning for Pedestrian and Cyclist Detection

A subset of artificial intelligence and machine learning, deep learning (DL) was first introduced in the 1990s but has only recently been able to be used due to advancements and decline in costs of computational equipment (e.g., graphics processing units (GPUs)) and efficient training algorithms [44,62]. In particular, the Convolutional Neural Networks (CNN) algorithms have been used in the field of computer vision and image analysis [59] for object detection [51], image classification [7] and face recognition [63]. CNN approaches have been considered state-of-the-art in this field of computer vision.

Convolutional Neural Networks (CNNs) are a type of DL technique that has high performance in many fields as object recognition and classification. These objects can include faces and handwritten numerals and letters. The robustness of CNNs stems from the fact that they are able to extract information from raw-pixel content and learn features automatically [44]. It does this by performing various operations, typically some combination of filtering, pooling and non-linear activation. One benefit of using CNNs for feature extraction, when compared to hand-crafted methods, is that CNNs learn features from the images without explicit programming.

Since 2012, new approaches based on DL techniques have developed for pedestrian detection such as AlexNet [1], a CNN technique developed by Alex Krizhevsky and named after the developer [64]. AlexNet was trained used in an ImageNet dataset. For ImageNet, the custom is to report two error rates; top-1 (full testset) and top-5 (fraction of testset). AlexNet error rates were 37.5% and 17.0% for top-1 and top-5 respectively. Prior to AlexNet's results, the best performance in terms of error rates were 47.1% and 28.2%. These results aided in the designing of hardware to improve the performance of CNNs for an increased accuracy in detection as well as the affordability of training of the CNNs.

DL uses multiple layers, which are able to extract features, such as edges or patterns in images and use these features to classify an object. In this way, deep neural networks such as CNNs are used for feature learning to recognise objects such as pedestrians [59,65–67]. Feed-forward neural networks comprise of a series of computational nodes known as neurons that are interconnected for information processing. This is also known as multi-layer perceptron (MLP). The nodes form layers that are interconnected through parameter values called weights. The neuron functions as a logistic regression classifier. The neurons use non-linear operations to transform input data and create a decision boundary in which the data can be linearly separable. An illustration of a single perceptron can be found in Figure 3. Multiple layers of these perceptrons create an MLP or neural network (Figure 4). The neural network in Figure 4 is a fully connected network. This means that each neuron receives an input from each neuron from the previous layer. For CNN, convolution layers exist within the hidden layers to perform the convolutional computations.

DL aims to detect objects in a single/multiple frames similar to how humans detect and interact with objects [44]. However, the detection of pedestrians and cyclists has been a major challenge in computer vision. With the recent software and hardware advancements, there has been a real progression in this field. There are many survey papers for pedestrian detection [46,68–74] and tracking systems, including the sensor technology and processing techniques. The use of a monocular camera for capturing images of pedestrians was used in [75]. A review of techniques of pedestrian detection techniques is compared, including some DL techniques, namely, the Convolutional Neural Network (CNN) in [43]. However, with the recent adoption of Deep Learning (DL) techniques,

state-of-the-art survey for pedestrian detection and tracking using these DL techniques should be conducted [23].

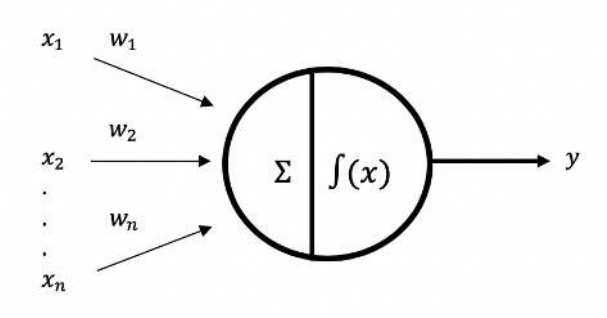


Figure 3. Single perceptron.

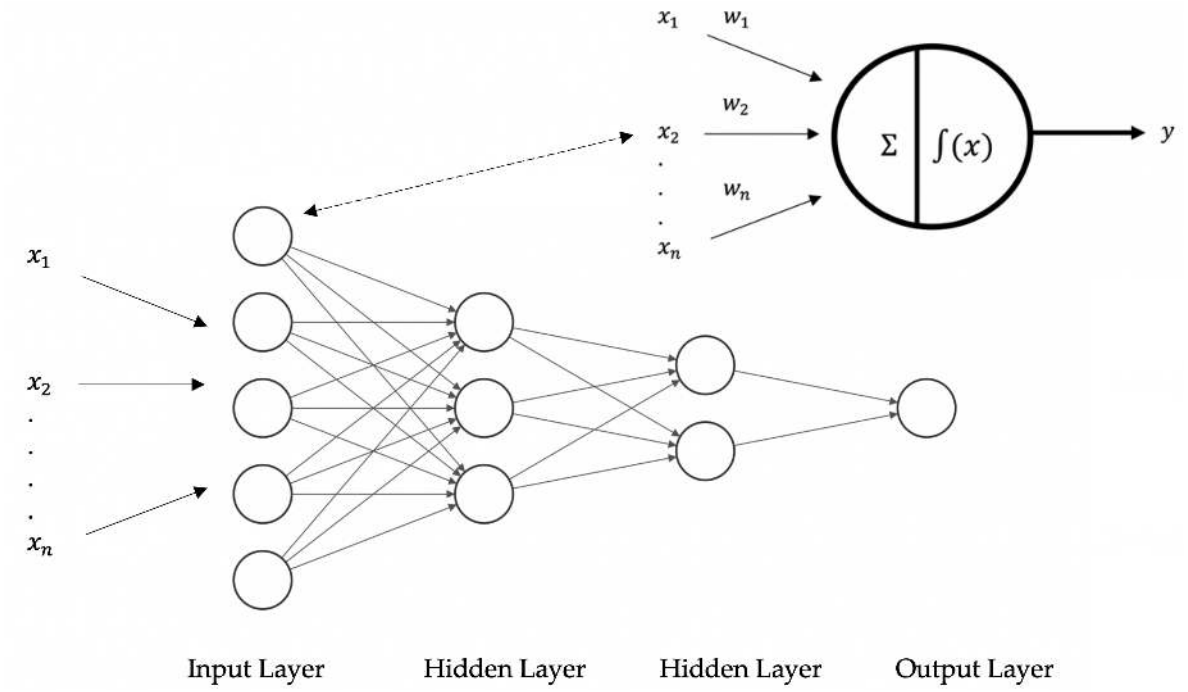


Figure 4. The architecture of a multilayer perceptron.

With the introduction of DL techniques (mainly CNN-based), deep network architectures are able to propose ROIs and extract the features for classification with a single step [23,64,76,77]. By this way, the need for traditional region proposal feature extraction techniques becomes obsolete. As deep networks can achieve a higher level of abstraction than traditional methods, higher accuracy and faster run-time can be achieved by deep network-based detectors [44]. This is one of the benefits of using DL for object detection. However, the training of these deep networks structure requires longer to build as deep networks require large annotated datasets for training. DL-based object detection has yielded encouraging results in the field of pedestrian detection and general object detection [50,64,78,79] (will be discussed in later sections).

Convolutional Neural Network

Prior to current state-of-the-art neural networks being introduced, basic neural networks (such as in Figure 4) would sometimes find it difficult to extract useful features from raw data from sensors. To find significant features, hand-crafted methods were used [65,66] (as discussed in Section 3). To overcome this and increase the performance of the neural network, Convolutional Neural Networks (CNNs) were implemented [59,65–67] (see Figure 5). CNNs are based on the feed-forward neural

network (where the output of a neuron would be the input of another set of neurons in the preceding layer). CNNs use convolutional operations to extract features from the input data (e.g., images, videos); with each layer using a kernel (filter) to extract input features. The activation value of the neurons in the layers represents the filtered input data. Different regions of the input are processed using convolutional operations to detect patterns in the data. Feature maps are then generated after the convolutional operation is performed across the entire input data [65]. The feature map is a representation of the activation of different parts in the image. It is used to set the parametrisation of the weights and biases of the layers, allowing the learning of features. Max pooling is typically used after the convolution to reduce the size of the input. This reduces the computation requirement as the parameters of the input are reduced. This also aids in over-fitting.

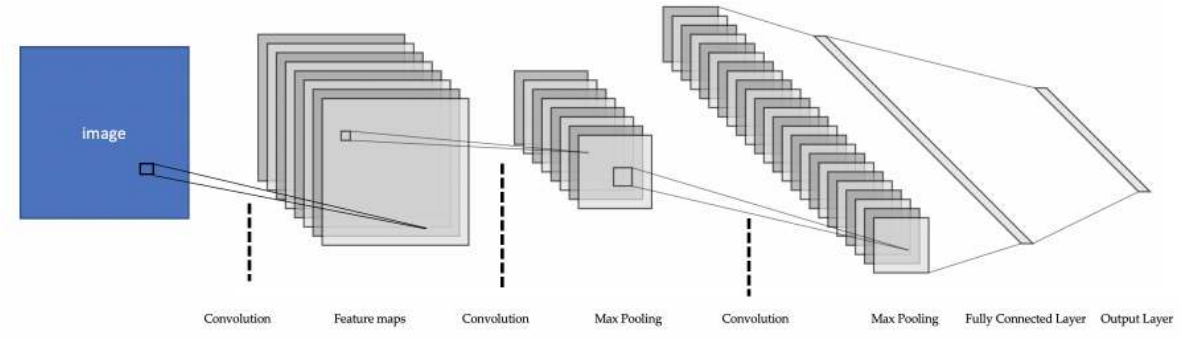


Figure 5. CNN architecture.

The convolutional operator that is used is dependent on the type of input data. 2D kernels (i.e., filters) are used for 2D temporal sequences (e.g., videos) and 1D kernels are used for 1D temporal sequences. When the CNN's kernel is used in this way, they can be used as classifiers [65]. With several layers, CNNs are able to represent data in a hierarchical fashion. As the layers become deeper, the input data is represented in a more abstract manner, something hand-crafted feature extractors would find very difficult or impossible to achieve. This has allowed CNNs to become more of a standard practice in many fields, such as computer vision (e.g., object detection) [64,80,81] and speech recognition [82].

The network can automatically learn to extract useful information (i.e., features) from images/frames. As the CNN is a DL technique, there will be numerous neurons and layers. Each layer will learn different levels of abstraction. The first few layers learn lower level features such as edges, curves or patterns. The deep layers will attempt to combine the features to identify objects in the frame [44]. The classifying layer typically consists of a number neurons. The number of neurons is dependent on the number of desired outputs (i.e., number of classes). For example, the classes could be pedestrian, cyclist or car, which means three classes are required. The higher the output value for one of these classifier neurons, the higher the chance that a pedestrian or cyclist is successfully detected. It is important to understand that this gives the deep network the ability to learn features without explicit programming. The learned information is stored within adjustable parameters of the network known as weights and biases. To train the network to learn features, a dataset is used. The dataset will feed numerous number of images that include the object that is to be detected. In this way, features are extracted and learned by the network. However, as the network learns based on only the dataset provided, it can be limited. Therefore, to design a more robust and accurate CNN, a very large annotated dataset is required.

6. Deep Learning Architectures for Pedestrian and Cyclist Detection

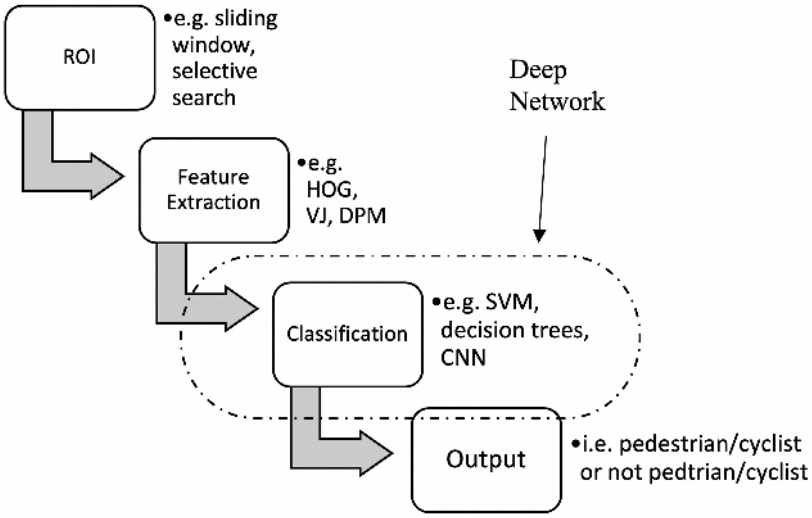
DL approaches for pedestrian and cyclist detection can be one of the following two categories: a two stages (region proposal approach) detector or a single stage (non-region proposal approach) detector. The single stage detector aims to remove the need for traditional region proposal feature extraction by processing these steps within a single network. The single stage detector can be simpler

to train, with a higher computational efficiency [5]. In this approach, a proposal of regions is first completed and then the deep network conducts the classification.

With the progress of DL and its development and success in pedestrian detection, detection accuracy has improved. The DL techniques used for pedestrian detection can include region proposal as part of the system. Some of the region proposal-based techniques include Region-CNN (R-CNN) [79], Regional-Fast Convolutional Network (R-FCN) [83] and Faster R-CNN [59]. Non-region proposal-based techniques include Single Shot Detector (SSD) [84–86] and You Only Look Once (YOLO) [87]. All of these pedestrian detection techniques are based on CNN, which has become the standard for pedestrian detection. For the task of classification, detection techniques can be placed into one of these families: DPM variants, decision forests and deep neural networks [47,48]. These techniques can also be applied for cyclist detection as they are visibly similar to pedestrians [25].

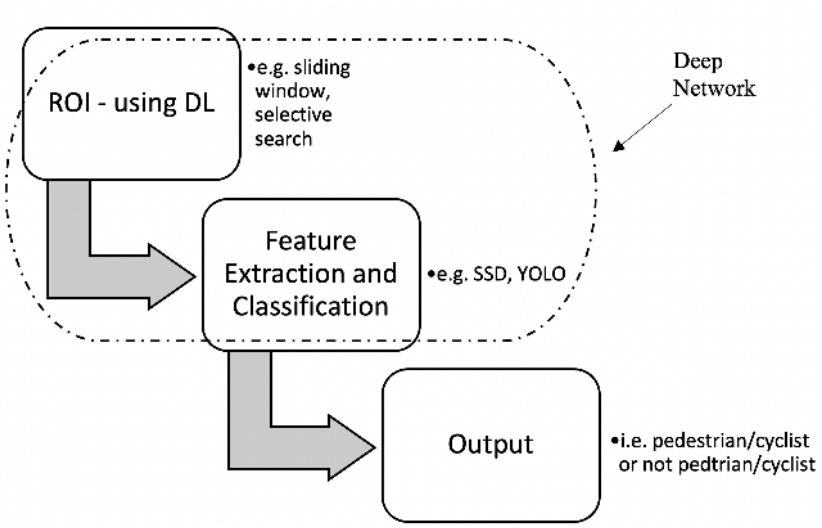
CNN is a popular technique for object classification in pedestrian detection systems [59,65]. An in-depth review of DL techniques is provided in later sections. A region proposal technique (such as the hand-crafted techniques described in the previous section) can be used alongside CNNs for object detection [59,65–67,79]. The region proposal technique is used to suggest where an object may exist in the image. The proposed regions are fed into a classifier (e.g., CNN) to determine the class of the object. Studies of the use of deep networks for pedestrian detection applications can be found in [88–92].

There are also non-region proposal-based DL techniques [59,79,83]. Figure 6 depicts the difference between the two types of architecture for object detection. Figure 6a is a detector based on traditional region proposal and feature extraction techniques, where only the classifier is the deep network. Figure 6b represents a deep network that is able to complete region proposal and feature extraction as well as classification in a single step. This is known as a single step detector. In 2009 the Caltech dataset was introduced for benchmarking the various techniques for pedestrian detection. The ConvNet (a CNN-based approach) was introduced in 2013 with competitive results when compared to previous pedestrian detection techniques mentioned [43].



(**a**)

Figure 6. *Cont.*



(**b**)

Figure 6. Architectures for CNNs for feature extraction and classification based on traditional methods for (**a**) Non-Region Proposal-based Detectors and DL methods for (**b**) Region Proposal-based Detectors.

6.1. Two-Stage Detectors

Region proposal-based CNNs (i.e., R-CNN, Fast R-CNN) has provided positive results for general object detection. An example of such a system is the use of selective search [60] in [79] for generating ROIs. The accuracy of this type of network is dependent on the region proposal technique that is applied as these ROIs are used for classification. Approaches have been made to improve the speed of two-stage detectors as in [51], where feature maps are generated when the deep network extracts features from the ROIs [59]. These techniques have since been adopted and variations of the techniques have been applied with encouraging results [5].

For example, in [93] the Average Precision (AP) achieved a higher score than the KITTI (Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago) evaluation [94]. The resulting AP increased by 9% to 16.7%. With the development of Fast R-CNN [79] and Faster R-CNN, computational speed has increased. Fast R-CNN is based on R-CNN, also designed by [79]. R-CNN uses selective search (a region proposal technique) to generate 2000 region proposals rather than some large number of region proposals. These proposals are fed into a CNN to feature extraction and then an SVM for classification. There were a few issues with this approach. Namely, classifying 2000 region proposals still takes a large amount of time. Also, this technique could not be implemented in real time as it took 47s per image. To overcome these issues, [95] (the same author who proposed R-CNN) introduced Fast R-CNN. A similar approach to the R-CNN, however, this time the image is fed into a CNN to generate a feature map. This feature map is used to identify region proposals (i.e., ROIs). This was faster than the R-CNN technique as the convolution is completed once per image rather than for 2000 region proposals, which could have more than 2000 region proposals. Fast R-CNN was found to be approximately 2 magnitudes faster than techniques based on R-CNN [95].

Faster R-CNN, proposed in [59], which lets the network learn region proposals, rather than use selective search, as selective search can be a time consuming process. Based on the Fast R-CNN technique, images are fed into a CNN to generate feature maps. However, instead of using selective search for identifying region proposals, a sub-network is used to predict region proposals. This sub-network, termed Region Proposal Network (RPN), learns region proposals using DL algorithms. RPN provided a mean average precision (MAP) of 75.9%, which is approximately 10.2% better compared to selective search results on the VOC (visible Object Classes) 2012 dataset [59].

However, these networks have also shown that computation of region proposal for object detection has a bottleneck as they are dependent on traditional techniques of region proposal.

Region-based CNNs share convolutions across proposals to reduce computational costs as in [79,96]. However, with the Fast R-CNN region proposal could be a bottleneck for the advancement for real-time detectors. To overcome this issue, a Region Proposal Network (RPN), which shares convolutional features with the detection network was proposed in [59]. This allows for region proposals that are almost computationally cost-free. The RPN is a CNN that functions by predicting object bounds (region proposals) and scores for those bounds simultaneously. This provides the detector with high-quality region proposals. This design performed at near real-time frame rates, improving the quality and object detection accuracy for general DL-based object detection.

For example, the Faster R-CNN, a two-stage detector, comprises of a region proposal network (RPN) and a classification sub-network. The RPN uses DL techniques to learn features in images, allowing it detect potential region proposals. These region proposals are then fed into a classifier to determine the class of the object. The Faster R-CNN has had state-of-the-art performance results on datasets, such as the PASCAL-VOC and Caltech datasets. Most notably, Fast R-CNN [25,40] and Faster R-CNN [25,40,48,79] have been used for pedestrian detection [59,97]. For example, in [40], there was an approximately 23% error reduction using a type of R-CNN approach when compared to some state-of-the-art techniques for pedestrian detection. These types of results illustrate the effectiveness of these techniques.

6.2. Single Stage Detectors

As promising as two-stage detector may be, for them to be able to process sizeable proposals, the computation is typically heavy in the second stage (i.e., classifier stage). So, Single Stage Detectors (SSDs) have been proposed that do not rely on region proposal in the hope that they would increase the speed of the system. SSD, such as You Only Look Once (YOLO), are designed in such a way that a single network predicts region proposals as well as the class of those region proposals [84]. This design saves a significant amount of computational time, allowing it to perform 3× faster than the state-of-the-art Fast R-CNN while achieving higher accuracy in [59]. Approaches for using deep networks for region proposal can be found in [80,98,99].

The two-stage techniques are implemented to increase the accuracy and speed of the network (when compared to R-CNN approaches), whereas, the single stage techniques focus on the overall speed of the system, allowing them to be better suited for real-time applications [19]. A comparison of the DL architectures can be found in [100] and summarised in Table 1. edited table below

Table 1. Detector Types.

Type	Advantages	Trade-Offs
Two-stage	Increased accuracy Information rich	Slower speeds Complex computation
Single stage	Higher speeds	Information loss Large number of false positives

7. Sensors Fusion Techniques Using Deep Learning

Even with recent development and advancement made in computer vision for pedestrian and cyclist detection, there are several challenges that need to be addressed [101–104]. One of the biggest problems is accuracy; which is affected by cluttered backgrounds, environmental conditions, occlusions [105], and poor visibility [104,106].

The environment around the autonomous vehicle is perceived through sensors. These sensors collect the environmental information that is then used for detecting any pedestrians or cyclists. Sensors can be classified as either active or passive. Active sensors typically require a device to be

attached to the object that is to be detected and tracked. Although active sensors provide simple processing, they have been typically applied to controlled environments [23]. For uncontrolled environments, passive sensors are more suitable as they use natural-based signal sources (e.g., natural light, thermal readings). Therefore there is no requirement for a device to be attached to the object, making it less intrusive than active sensors. Passive sensors would be more effective for autonomous vehicle applications as the environment in which they operate will be uncontrolled and attaching tracking devices would not be feasible. Some examples of the types of sensors employed in vehicles for environmental sensing are visible cameras, thermal cameras, LiDAR and RADAR. Implementation of these sensors is also described in Table 2. The primary focus of this review will encompass visible and thermal sensors as it has become apparent that further research surrounding these sensors are required.

Table 2. Sensors for detection.

Study	Sensor Type [1]	Overview	Purpose [2]	Performance	Evaluation
[107]	VS & IR in pairs (4 in total)	Flow from the cameras is processed independently and then fused. This method provides a list of detected pedestrians.	PD	Even when not visible by the visible cameras, pedestrians were still detected. Works even when pedestrians are occluded. This approach was able to detect more than 95% of pedestrians at 45 m and more than 80% at 75 m.	The camera system consisted of 2 colour cameras and 2 far infra -red cameras. It was evaluated over 5000 images.
[108]	IR and laser filter	Kalman filters in parallel to handle fusion of sensors for detection and localisation	PD +T	The multi-sensor approach uses an IR camera for detection and laser for tracking. The technique aids in providing a precise location of the pedestrian(s). Proved to work well even when pedestrians are overlapping.	The sensor system was implemented to obtain real-time results. These results were discussed in the study.
[109]	VS, radar and LiDAR	Fusion at detection level, reducing the number of false detections	T	The IR sensor was used as it is not affected interference from visible light. It is also cheap and easy to implement and provides a long detection range. The techniques were able to increase the accuracy and acceleration of the tracking.	Datasets for evaluation were generated using the CRF (Fiat Research Center) demonstrator for various driving scenarios that can be encountered in real-time.

Table 2. *Cont.*

Study	Sensor Type [1]	Overview	Purpose [2]	Performance	Evaluation
[110]	VS, IR, LiDAR and RADAR	Improve standard ADAS with processing units. Combining the various sensors as part of the ADAS can provide an improved detection system	PD +T	Improved environmental detection. RADAR and LiDAR provide precise distance measurements and is not influenced by weather or low illumination conditions. Unlike the VS, camera, the resolution of the RADAR and LiDAR sensors is affected by elevation. However, the VS camera is affected by illumination and weather, while RADAR and LiDAR are not.	The study discussed the benefits of sensor fusion. This study was an informative piece rather than a implementation of a proposal.
[111]	RADAR	FFT processing radar for the distinction of moving targets from background	PD +T	Information successfully extracted for slow moving humans from the background. During the evaluation and testing a moving human was successfully detected at 1.76m with a velocity of -4.39 km/h in a crowded scene. The proposed method was able to provide a clearer than typical tracking systems based on RADAR.	The detection algorithm was tested using real-time data and a 24 GHz radar transceiver.
[52]	VS	Multi-view detector for different viewpoints and SVM classifier	CD +T	Successful tracking, even with changes in orientation of cyclists. HOG descriptor for feature extraction and an SVM for classification. It was an effective method, however, it has not taken bicycle kinematics into account.	Various datasets, particularly the INRIA dataset, were used for testing the proposed method as well as a custom dataset created from collecting online images.
[54]	VS	More effective feature extractor: HOG-LP	CD	The proposed HOG-LP technique was intended to overcome the shortcomings of the original HOG descriptor. The method was designed for detecting cyclists that were crossing the road. It was able to achieve a detection rate of 93.9%, with a false positive rate of only 0.3%. However, it was stated that it would be not suitable for real-time application as its speed was not fast enough.	At the time of this study, there was no public dataset for cyclist detection. Therefore, data was collected to created a cyclist dataset. 1000 positive samples were collected. 400 samples were used for training and the remaining 600 samples along with an additional 3000 negative samples were used for evaluation.

[1] VS-visible Data, IR-infrared Data. [2] PD-Pedestrian Detection, CD-Cyclist Detection, T-Tracking.

7.1. Visible-Based Sensors

For the detection of pedestrians and cyclists, visible sensors are often used as they are able to capture high-resolution images [107,108]. The images provide useful information for detection and classification, such as colour and texture. Cameras also typically provide more information than active

sensors [25]. Visible cameras have been applied to multiple tasks for vehicles, such as lane detection, distance detection from other vehicles and traffic sign detection.

In terms of 2D and 3D visible cameras, a 2D video would provide enough information to perform object detection. It could potentially even allow for tracking using a bi-dimensional approach [23,112]. However, 2D cameras lose a large amount of information when used with a bi-dimensional approach [23], and therefore may not be suitable as that lost data may have held useful scene information. The 3D cameras create a virtual environment in which the pedestrian coordinates can be represented in 3D space. Unlike a 2D system, a 3D system uses a stereo camera (i.e., multiple lenses), allowing the camera to capture 3D views based on multiple points of views.

Even though vision-based detection has been extensively researched in recent years, there are still issues and challenges while the detection system is in operations [25,106]. These are caused by the appearance of the pedestrians/cyclists due to occlusion, pose, crowded scenes and clothing. Cyclist detection, however, can be more difficult than pedestrian detection as cyclists can have a greater number of possible orientations. To overcome this challenge, combining visible cameras with other sensors could be beneficial.

7.2. Thermal-Based Sensors

Even though visible sensors have difficulty functioning when there is a low level of light (i.e., night-time, bad weather), they are the most commonly used sensors for pedestrian and cyclist detection applications [106]. To overcome the shortcomings of visible sensors, thermal cameras could be used in conjunction with visible cameras as, unlike visible cameras, thermal cameras are not significantly affected by ambient lighting [113]. A type of 3D system that could use both visible and thermal information to provide more accurate detection and tracking, known as an RGB-D was proposed in [114,115]. RGB data provides textural and appearance information of the object being detected, while the depth data (e.g., thermal data) can provide additional information of the shape of the object [106]. This approach was implemented in [106] by fusing an RGB camera with a depth camera, which detects heat signatures using a thermal sensor [116,117].

There are two types of thermal sensors that can be used for pedestrian and cyclist detection applications, the Near-IR (infrared) camera and the Far-IR (also known as thermal) camera. Near-IR has wavelength of 0.75–1.3 μm and Far-IR cameras have a wavelength of 7.5–13 μm. Pedestrians and cyclists would appear more visible to the thermal cameras than the near-IR, as the pedestrian/cyclist body heat radiates in the long-wavelength (approximately 9.3 μm), making the thermal camera ideal [108,110,118]. The value of the radiation emitted from a human is not particularly affected by other illuminations in the environment (e.g., street-lamps, artificial lighting). This can improve the accuracy of the detector, as demonstrated in [106,119–123]. In [123], a commercial visible camera with a resolution 640x480 was used with a thermal camera. Testing was completed during various times of the day and weather conditions (i.e., morning, night, afternoon, rain etc.). In one of the tests, using both visible and thermal cameras, the accuracy of the system was 98.13%. When comparing this result using the visible and thermal cameras separately, the accuracy was 72.11% and 95.91% respectively.

Comparison of the benefits and drawbacks of visible and thermal cameras, RADAR and LiDAR can be found in Table 3 and Figure 7. Combining the sensor information can offset some of the inefficiencies of the individual sensors, such as a higher detection accuracy throughout the day, even in crowded scenarios.

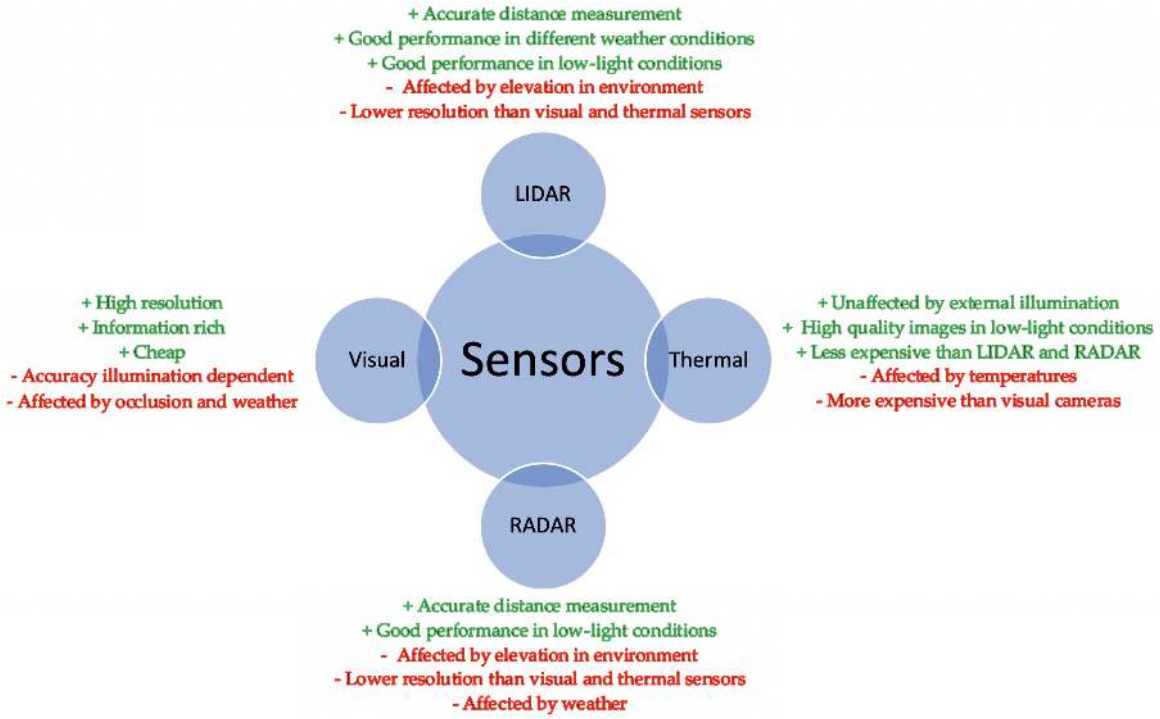


Figure 7. Sensor Comparison.

Table 3. Sensor Comparison Matrix.

Environmental conditions & costs	Visual	Thermal	RADAR	LiDAR
Resolution	●	●	●	●
Illumination	●	●	●	●
Weather	●	●	●	●
Elevation	●	●	●	●
Temperature	●	●	●	●
Cost	●	●	●	●

●—good ●—fair ●—poor.

7.3. Sensor Fusion

Visible sensors are effective, however, they are less reliable in low-light situations [86]. The study in [86,97], suggests combining visible and thermal cameras together to increase the detection accuracy. It should be noted that thermal sensors are not very effective under high-temperature conditions and clothing can affect the pedestrian's or cyclist's thermal footprint [124].

There has been a large amount of research conducted surrounding the most reliable approach in using both colour information of the visible cameras and the thermal information of the thermal cameras [86,97,125–128]. These studies discuss the drawbacks of visible sensors due to their dependence on illumination and the benefits of adding thermal data would provide for increased accuracy. The KAIST dataset is largely used for multispectral pedestrian detection evaluation due to its large amount of high-quality images in both the visible and IR spectrum. In [86], fusion techniques using a CNN as the detector were discussed. Figure 8 compares evaluation based on the KAIST dataset and typical vision-based datasets (Caltech Diamler, etc.). It demonstrates that, although recent vision-based approaches are efficient during the day, their accuracy decreases at night.

Therefore, using multispectral information can aid in reducing this inaccuracy, especially at night-time. For further information of the studies used to generate the graph see [88,89,129–131] for VS data for daytime, [106,125,128,132] for VS data for night-time and [86,97,127,133–136] for multispectral data. Figure 8 demonstrates, that although there has been advancements in pedestrian/cyclist detection, improvements are still required to reduce the miss rate and increase the accuracy of detection systems. Multispectral information can be used to achieve higher accuracy, however, further investigation is required.

Vision-based pedestrian detection has been widely researched, which has provided over 60 methods evaluated on the Caltech dataset alone [25]. Despite sensor fusion techniques having been recently applied for pedestrian detection, there is not much research conducted for cyclist detection using the same multispectral approach. This is an important aspect to be considered as cyclists are among the VRUs that are affected by road traffic accidents.

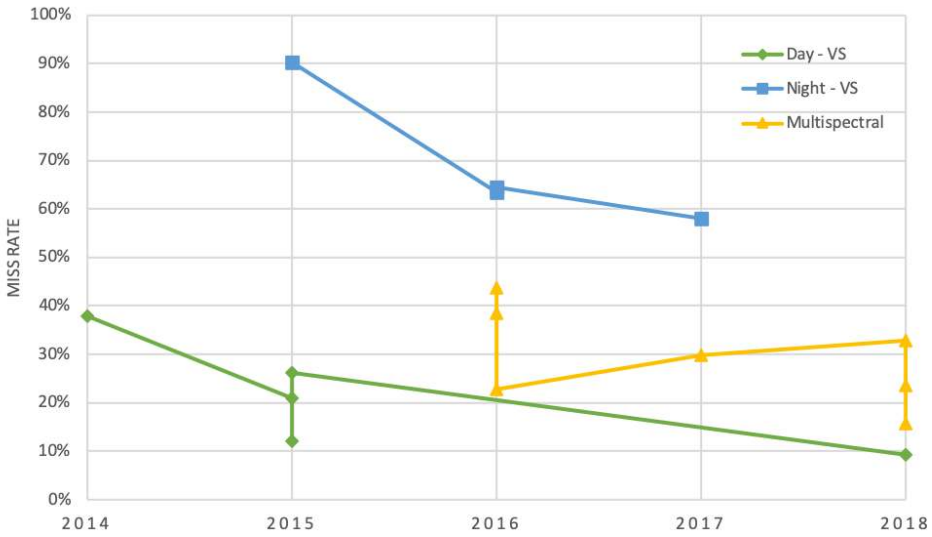


Figure 8. Miss rate of recent detection approaches by year using visual and multispectral data. Visual data for daytime and night-time and multispectral data is being represented by miss rate by year of publication. The lower this miss rate, the more effective the approach. The multispectral data in this case is focused on improving night-time detection efficiency. Multispectral data in these studies focused on using visual data with thermal data.

7.4. DL for Sensor Fusion

With the development of R-CNN, Fast R-CNN and Faster-CNN, CNNs have become a standard technique for detection and classification applications. As CNNs have provided positive results in the field of computer vision, studies have been undertaken for using CNNs with multispectral data for pedestrian detection.

The fusion of the sensor information can be either at pixel-level, early fusion (feature-level), late fusion (decision-level), or halfway fusion [86] (see Figure 9). To fuse the data at pixel-level, vision-based images are converted into the HIS (Hue-Intensity-Saturation) colour space. Thermal images are intensity images; therefore, the fusion of the thermal images and the visible images takes place in the intensity (I) component. The images are then reconstructed with the new I value. Some pixel-level fusion methods includes wavelet-based transform [137,138], curvelet transform [139] and Laplacian Pyramid fusion [140]. Pixel-level fusion is typically not used with DL-based approach for sensor fusion as the fusion takes place outside of the deep network. Therefore, early fusion, late fusion and halfway fusion are the typical architectures for DL-based sensor fusion. For feature-level (early fusion) visible and thermal images are combined together as a 4-channel input for the deep network (see Figure 9a). The network would then learn the relationships between the image sources [125]. In decision-level (late fusion), feature extraction takes place for both image sources into sub-networks

(Figure 9b). These features are then fused before being fed into network layers that classify the object. Halfway fusion involves feeding the colour and thermal data separately into the same network. The data is then fused inside the network (Figure 9c).

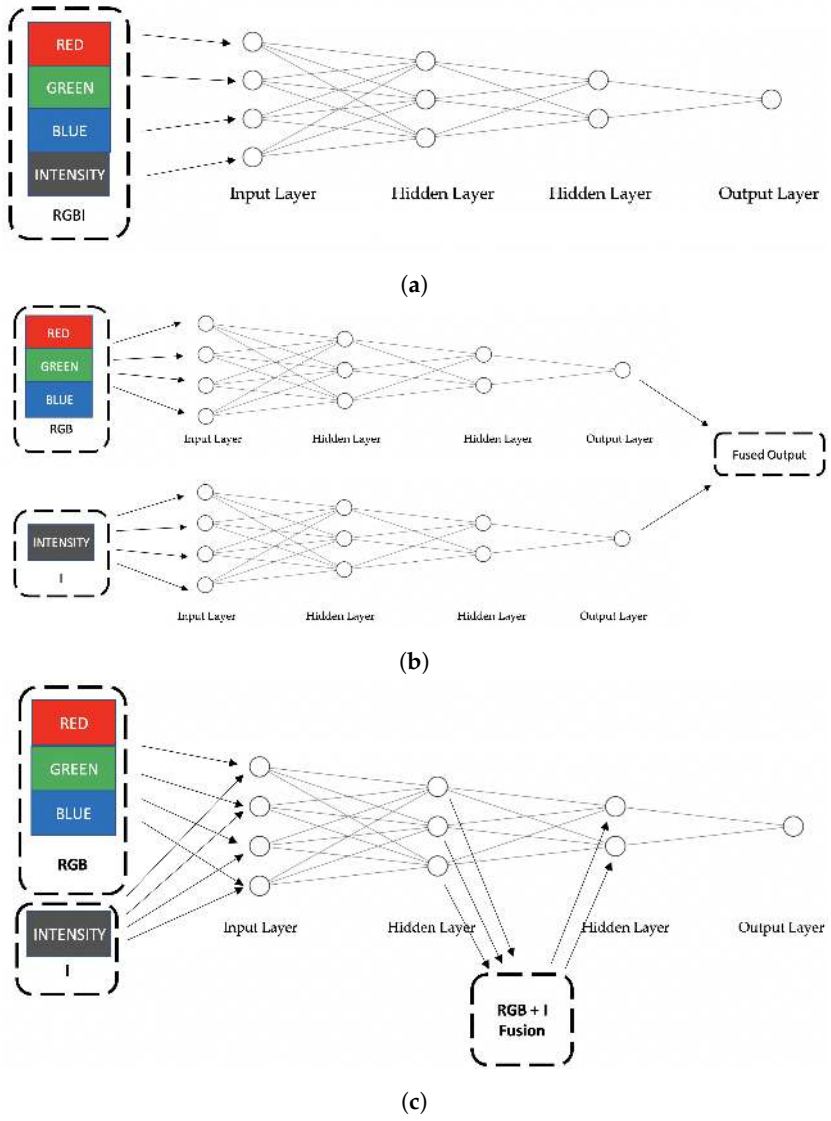


Figure 9. Sensor Fusion Techniques: (**a**) Early Fusion, (**b**) Late Fusion and (**c**) Halfway Fusion.

In [141], decision-level fusion techniques were used to combine the results of visible and thermal images for detection and tracking purposes. Hwang et al. [106] proposed a detector for pedestrians using a aggregated channel features (ACF) technique based on fused features of visible and thermal images. The benefits of using multispectral detection techniques are demonstrated by [97]. It was found that combining visible and thermal data produced the best results, however, during low-light conditions (e.g., night), the thermal sensor performed better on its own. Combination of the visible and thermal data actually performed worse at night with an increase of the Average Miss Rate (AMR) by 3%. Overall, data fusion decreases AMR by 5% compared to visible and thermal data used on their own during the daytime. This was unexpected as it was thought that thermal data would not add to the feature detection of visible images. Evaluation of the KAIST dataset produced competitive results (64.17% AMR) compared to state-of-the-art Caltech evaluation protocol (65.75% AMR) for pedestrian

detection. It should be noted that using more than a single sensor causes an overall increase in system complexity due to alignment and synchronisation of the cameras [97].

Fusion architectures were compared in [125], with halfway fusion proving to be the most effective, with a 3.5% lower miss rate (MR) than the other two architectures. Also, using a single form of sensor information (i.e., visible or thermal data) was shown to be worse than the Halfway fusion model with an increased MR by 11%. The evaluation was completed on the KAIST dataset. In [128], investigation for the optimal fusion technique for CNN-based pedestrian detection was undertaken. Fusion architectures were tested using a Faster R-CNN. Two types of fusion techniques were examined; feature-level and decision-level. Pixel-level techniques were not considered in this study. In another study [86], the performance of pixel-level fusion with early and late fusion techniques were considered. In [86], fusion architectures were implemented with an SSD. The results indicate that the Pixel-level Fusion does not perform well, but for early and late fusion, multispectral information can achieve better performance for pedestrian detection. Based on the KAIST dataset, using the wavelet transform for pixel-level fusion provided a lower Miss Rate (MR) for the early and late fusion of 9% and 5% respectively.

In [97], assessment of the gain of accuracy fused visible and thermal images was conducted. To achieve this, the results from visible images, thermal images and then a combination of the two was compared. An early fusion technique approach was used for the study. Evaluation and the results were compared using the KAIST [142] multispectral dataset. The study found that a combination of features of visible and thermal images produces a better detector than with visible images or thermal images alone in the daytime. This result was not what was expected as it was believed that thermal images would not improve the features in the visible images. There is also a slight improvement in the night-time for the combination of the images.

8. Datasets

Due to its real-world application and significance, pedestrian and cyclist detection has been a widely studied problem. The key challenges that associated with this field have been variations in pose, scaling and occlusion. These effects can be seen in major datasets, such as the Caltech dataset, where several pedestrians are affected by occlusion [25]. Pedestrians and cyclists are traditionally considered separately, which can lead to having the input image scanned multiple times to detect the two objects independently. This not only increases computational costs, but it can further cause detection errors where the pedestrian and cyclists are misclassified due to the similarity in appearance.

A dataset is used to train a deep network for pedestrian and cyclist detection. They can also be used for benchmarking and comparing the accuracy and performance of pedestrian detection techniques, as in Table 4. For deep network models, large annotated datasets are required to produce an accurate system [50]. The pedestrians, or any other object in the dataset, needs to have bounding boxes that are annotated. For general object detection, ImageNet has proven to be a sufficient dataset to train a CNN [44,143,144].

Table 4. Detection Methods Benchmark [3].

Method	Miss Rate	Dataset
Shapelet [72]	94%	Diamler [102]
FrtMine (Feature Mine) [145]	85%	Caltech Japan [11]
Pls (Partial Least Squares) [146]	72%	Caltech Japan
VJ (Viola–Jones) [147]	72%	INRIA [32]
FPDW (Fastest Pedestrian Detector in the West) [148]	63%	TUD-Brussels [149]
ChnFtrs (Channel Features) [150]	60%	TUD-Brussels
MultiFtr+CSS (Multiple Feature+Color Self Symmetry Information) [151]	59%	TUD-Brussels
LatSVM-V1 (Latent SVM) [46]	58%	Diamler
MultiFtr [152]	57%	Diamler
MultiFtr+Motion [151]	55%	Caltech [11]
HikSVM (Histogram Intersection Kernel SVM) [153]	55%	Diamler
LatSvm-V2 [35]	51%	ETH [154]

Table 4. *Cont.*

Method	Miss Rate	Dataset
HogLbp (HOG Local Binary Pattern) [34]	49%	Diamler
HOG [32]	45%	INRIA
JointDeep [155]	45%	ETH
HogLbp	39%	INRIA
MultiFtr+CSS	39%	Diamler
LatSVM-V2 [49]	38%	Diamler
MultiFtr	36%	INRIA
ConvNet (Convolutional Networks) [156]	33%	Diamler
MultiFtr+Motion	29%	Diamler
MLS (Macrofeature Layout Selection) [157]	28%	Diamler
WordChannels [158]	16%	INRIA
NAMC (Normalized Autobinomial Markov Channels) [159]	15%	INRIA
RandForest (Random Forest) [160]	15%	INRIA
SCCPriors (Symetric Cross-Channel Priors) [161]	15%	INRIA
Franken [162]	14%	INRIA
InformedHarr [163]	14%	INRIA
LDCF (Locally Decorralated Channel Features) [61]	14%	INRIA
Roerei [164]	14%	INRIA
SketchTokens [165]	13%	INRIA
SpatialPooling [166]	11%	INRIA
RPN+BF (Region Proposal Network+Boosted Forest) [167]	10%	Caltech
MS-CNN (Multi-Scale CNN) [90]	10%	Caltech
SA-FastRCNN (Scale Aware-FastRCNN) [131]	10%	Caltech
UDN+ (Unified Deep Network) [168]	10%	Caltech
Adaptive Faster R-CNN [169]	9%	Caltech
F-DNN+SS (Fused Deep Neural Network+Selective Search) [170]	8%	Caltech
F-DNN2+SS [170]	8%	Caltech
TLL-TFA (Topological Line Localization-Temporal Feature Aggregation) [171]	7%	Caltech

[3] Effectiveness of different methods evaluated on datasets for pedestrian detection. These results are from the Caltech Pedestrian Detection Benchmark and are arranged in a log-average miss rate (MR), where the lower the value, the more effective the method. The evaluation for the benchmark can be found in [11,172].

Datasets are required for training the network to learn features for classification of pedestrians and cyclists. There exists several datasets for pedestrians, however, cyclists datasets are limited.

8.1. Pedestrian Datasets

Although vision-based approaches cannot collect the same level of information at night-day or low-light conditions, most of the detectors are based on colour images. This is due in part to the fact that many of datasets that are used for benchmarking are of colour images [106]. Some of the most commonly used datasets for evaluating pedestrian detection techniques are Caltech-USA [173], KITTI [104], ETH [174] , TUD-Brussels [175]. These datasets are vision-based datasets [43]. Of these datasets, Caltech and KITTI are the major benchmarks for pedestrian detection as they are large datasets with pedestrians in scenes that pose challenges for detection due to crowded scenes, occlusion, etc. [25].

The KAIST [142], a multispectral dataset, combines data collected using visible and thermal cameras. The KAIST dataset aims to improve the training of deep networks for detection, as most of the large-scale (i.e., Caltech, KITTI, etc.) use only RGB-based image. The issue with using only colour images is that it is assumed that the autonomous vehicle would be functioning in a well-lit condition, which is not always the case in practice. Therefore, thermal images can be used when the conditions do not allow visible images to function as designed. As KITTI has been so widely used, [142] was influenced to provide a similar quality dataset. They also used KITTI as a ground truth and valuation criteria.

8.2. Cyclist Datasets

There is significantly more dataset for pedestrians than there is for cyclists. A public dataset for cyclists was introduced by [48] known as the Tsinghua-Daimler Cyclist Benchmark. Prior to this, there was no challenging dataset for cyclists. Although there was an object detection benchmark as part of the KITTI dataset, however, it contained less than 2000 cyclist instances. This could be seen as insufficient for training a detector and evaluation (i.e., testing). Cyclists have been often disregarded due to their similarities with pedestrians. However, cyclists can be just as vulnerable as pedestrians, and therefore Tsinghua-Daimler Cyclist dataset was introduced. Based on the Tsinghua-Daimler Cyclist dataset, a new dataset was presented in [25,48], which contains both pedestrians and cyclists. They added a pedestrian dataset into the cyclist dataset as no dataset for both pedestrians and cyclists exist at this time. A comparison of the datasets mentioned above can be found in Table 5. It should be noted that some other pedestrian datasets, such as KAIST, do contain cyclists but not enough to train and evaluate a network.

Table 5. Datasets for Pedestrians and Cyclists.

Object Type	Visible	Multispectral	Intent Estimation
Pedestrian	Caltech [173] KITTI [176] CityScapes [177] ETH [69] TUD-Brussels [175] PASCAL-VOC [179]	KAIST [106]	Daimler [178]
Cyclist	Tsinghua-Daimler [48] KITTI [176]	FLIR [180]	———

9. Deep Learning for Intention Estimation

Intent estimation is the latter part of the detection and intent estimation system (Figure 1). Avoiding incidents involving autonomous vehicles and VRUs is a critical aspect of the fully automated vehicles [14,181,182]. Even with success in pedestrian detection using deep networks [5], it is not enough to simply identify the pedestrian or cyclist. The autonomous vehicle must also predict if there is a chance of any harm that may befall the pedestrian/cyclist due to the autonomous vehicle action or inaction [181]. Predicting the motion and future behaviour of VRUs can aid in improving safety for autonomous vehicles and other road users, and, therefore, can be considered as a critical part of self-driving vehicles [183]. Once detection of the pedestrian/cyclist is completed, it must be considered whether the pedestrian may be in danger given their location with respect to the current distance, motion and path of the autonomous vehicle. Even when detection and localisation of the pedestrian are achieved, pose estimation and tracking must be considered as it can allow the vehicle to take actions or manoeuvres to prevent accidents if the pedestrian's or cyclist's intentions are considered to cross the path of the vehicle [184]. However, this can be difficult, especially for long-term predictions. A major cause for this is the agile nature of a human, who can very quickly change speed and direction, and may not allow the vehicle to have sufficient time to react. This limits the reliability of the prediction system as seen in [185]. Reduced accuracy in prediction could mean that the vehicle could misinterpret or not react in time to a pedestrian's or cyclist's sudden movements. In [183], minimisation of the false detections is described for improved accuracy for long-term predictions. DL has been used for intention estimation [186,187] with promising results as will be discussed in this section. Some of the literature for intent estimation can be found in [178,183,188–198] which is summarised in Table 6.

Table 6. Intent estimation approaches.

Study	Approach [4] & Sensor	Overview	Technique [5]	Performance	Evaluation [6]
[178]	DMM, VS-based	Comparative study of recursive Bayesian filters for pedestrian path prediction. The purpose of the paper was to explore the accuracy and benefits of single/multiple models with EKF.	EKF and IMM using single/multiple models.	The models were applied for four pedestrian motions; crossing, stopping, bending and starting. Position measurements were obtained using a vision-based pedestrian detector The results showed that the single and multiple models had very similar performance in terms of position estimation. This could be due to a high sampling rate and low measurement error. However, for path prediction, the IMM outperformed the single model-based approach and could improve position estimation of up to 30 cm.	Image sequences were recorded using a stereo camera system. The dataset contained 12,485 images containing pedestrians. A state-of-the-art HOG-SVM detector was used for detecting the pedestrians in the images.
[199]	DMM, VS-based	Prediction of pedestrian locations and pose to classify intentions up to 1 s ahead. Location and pose prediction of pedestrians for intention classification.	B-GPDM and naïve -Bayes classifiers	Intention prediction up to 1s ahead of time using this technique. The approach added in reducing the number of misclassifications as well as avoiding continuous action changes of the pedestrian, such as walking and stopping. The average mean error was 29.47 cm for stopping, starting and walking trajectories.	Evaluated based on the CMU dataset, which contains 63,508 poses based on 129 sequences. The CMU dataset is vision-based.
[200]	PBM, VS-based	The purpose of the study was to estimate the probability distribution of the future positions of a pedestrian based on plan planning approaches.	Particle filter	The pedestrian's future destination is estimated so that the position of the pedestrian becomes a path planning problem, taking environmental conditions into account. Unlike DMMs, other models, such as dynamic states or behaviours models, are not required. Instead these models are solved implicitly. Dynamic states can include passing cars and cyclists. This method has provided higher levels of performance when compared to DMMs.	The method was tested using the dataset presented in [201]. This dataset is vision-based. There is 2113 frames of pedestrians stopping and 2436 walking.

Table 6. *Cont.*

Study	Approach [4] & Sensor	Overview	Technique [5]	Performance	Evaluation [6]
[202]	DL, VS-based	Long-term intent prediction of VRUs based on motion trajectories. Treated as time series problem.	RNN & LSTM architecture	When evaluated and compared to state-of -the-art baselines, the approach provided improved results in terms of overall mean lateral position error. In one of the evaluations, there was an improved result of up to 85% for the standing sequence. However the evaluation for the stopping sequence produced an increase of mean position error rate.	The dataset used for testing the proposal was the Daimler pedestrian path prediction benchmark dataset as presented in [178]. It contained 68 sequences and based on a stereo camera system.
[181]	DL, VS-based	Vision-based pedestrian intention estimation. CNN is used to detect and provide skeleton information.	CNN and SVM	The CNN is able to extract high-level features. These features are then processed using an SVM. The high-level features provide more information about pedestrian actions than low-level features, such as HOG and Histogram of Optical Flow (HOF). Unlike typical detectors, a monocular camera was implemented for this method. This method was able to predict the intention of a pedestrian 750 ms before a pedestrian will cross while walking and 250 ms after a pedestrian moves from the bent forward position. In addition, 187 ms when entering the road from a stand still position. Intention estimation is still difficult, especially at a distance or crowded situations.	A stereo camera system was used for testing. The author mentions that others have used LiDAR as well. The Daimler pedestrian path prediction benchmark dataset was used.

[4] DMM-Dynamical Motion Modelling, PBM-Planning-Based Models, DL-Deep Learning, VS-vision. [5] EKF-Extended Kalman Filter, IMM-KF-Interacting Multiple Model-Kalman Filter,B-GPDM-Balanced Gaussian Process Dynamical Models, RNN-Recurrent Neural Network, LSTM-Long Short-term Memory. [6] CMU-Carnegie Mellon University.

As illustrated in Table 6 and throughout this section, most research focuses on short-term path prediction based on visible cameras [185,189,203]. Although there has been studies undertaken for using thermal sensors to tracking pedestrians/cyclists (see [204,205]), there has not been as much focus on using thermal data for pedestrian/cyclist intent estimation. Therefore it may be useful to further investigate the improvements that may be brought from sensor fusion for VRU intent estimation in the same way as sensor fusion for VRU detection was discussed in Section 7.

As stated in [183,206], it was found that hand-crafted feature descriptors would not be able to provide the level of accuracy required to pedestrian detection. As DL approaches are able to extract features directly from the input data, making them effective in pedestrian/cyclist detection applications, this motivated the implementation of a deep network architecture for the purpose of pedestrian intention estimation in [183]. The network was coupled with a long-short-term-memory (LSTM) network to improve the accuracy of the system. LSTMs are used for learning in time series [207]. More on this work can be found in [183].

Another approach uses Cartesian coordinates with a Bayesian Network [185] and Gaussian process regression [208]. The Bayesian Network used multivariate Gaussian distributions for the relative position and velocity with respect to the vehicle for a predicted time. There has also been research undertaken using body features that can aid in predicting a pedestrian's future behaviour such as walking, standing, bending, jogging and running [188,209,210]. The head orientation of the pedestrian was considered in [183,203]. This feature allows for determining the awareness of the pedestrian situation and the approaching vehicle.

Before the advent of deep learning techniques, typically pedestrian trajectories using Kalman Filters or naïve movement models using human gait estimation and analysis of simple heuristics [211,212]. However, due to the improbability of proper adaptation and handling to changes in pedestrian movement, these techniques provided poor results in terms of predicting future pedestrian movements [1,178].

Other than DL approaches, there have been two typical approaches for predicting the future actions of a pedestrian [178]. One approach is based on dynamical motion modelling (DMM) [178,189,203]. This approach is able to predict the motion trajectories for different scenarios. However, the model assumes that all trajectories inhibit similar dynamics. This leads the model to have a lower accuracy when predicting long-term motion modelling for intent estimation. Planning-based models (PBMs) [213,214], has shown better results for long-term predictions. However, the model requires the final destination information of the VRU, which is a difficult task for a moving vehicle to infer. Although the DMM and PBM approaches have proven to be quite powerful, they are dependent on hand-crafted features (e.g., HOG, Haar-like, DPM, etc.). As discussed earlier, hand-crafted features for detection are limited. The same can be said for hand-crafted features for intent estimation. When the autonomous vehicle processes an unseen or complex situation, it may not be able to react properly as the hand-crafted features create a generalisation of certain situations. That is because feature selections and parameters when designed by an expert, rather than using real-time information collected from sensors. This causes these techniques to be rather restricted and under-perform in previously unseen situations. To solve this problem, [202] proposed a data-driven (i.e., DL) approach, where the initial motion trajectories [178] are used for long-term intent estimation, enabling for predicting future positions of a VRU up to 4 seconds ahead. A Recurrent Neural Network (RNN)+Long Short-Term Memory (LSTM) approach was adopted. The RNN+LSTM method is typically used for time series problem.

9.1. DMM

Dynamical motion modelling (DMM) is the general approach for the future location of pedestrians based on motion trajectory [202,215]. In [178], an Extended Kalman filter, a type of Bayesian filter, was used for short intent estimation (<2 s). Further details of this approach and some of the other approaches discussed in this section can be found in Table 6, a Dynamic Bayesian Network (DBN) was used for intention estimation of a pedestrian that is walking on the curb [203]. As part of the DBN, a Switching Linear Dynamical System (SLDS) was also implemented to predict the changes in the pedestrian's motion. In [199], the pedestrian intention was predicted using pose estimation based on dynamical models and behaviour classification using Balanced Gaussian Process Dynamical Models (B-GPDM) and naïve Bayes classifiers. Another approach uses a dynamic model with a HOG feature descriptor with Linear SVM detector [32]. It uses an Interacting Multiple Model based on Kalman Filters (IMM-KF) for future predictions of the pedestrian. However, simpler methods such as constant speed velocity models can provide comparable results to IMM-KF, which is a more complex technique [181]. The results were improved in [189] by using a Gaussian process dynamic model with probabilistic hierarchical trajectories. This approach uses the silhouette of the pedestrian and attempts to predict its future progress. The approach predicted the pedestrian's action with respect to the path of the vehicle. These methods require that common trajectories of pedestrians are learned and then these are classified. This means that the technique may not be reliable in previously unseen scenarios.

Therefore, as pedestrians and other VRUs are able to quickly change direction and motion, DMMs decrease in reliability for increased prediction lengths [200].

9.2. PBM

Unlike DMMs, planning-based models (PBMs) for pedestrian's future movements do not model the intentions of the targets explicitly [178]. Instead, they assume that the target (i.e., pedestrian) has the intention to reach a particular destination. For example, in [200], a model was proposed for long-term intention prediction of a pedestrian. In this model, the pedestrian's goal is to reach a location is predicted based on the estimation of the probability distribution of the future positions of the pedestrian. The model is based on a probabilistic path planning technique. A grid occupancy map is used to estimate the destination of the pedestrian based on the position and orientation of the pedestrian. The model was trained using supervised learning with pedestrian trajectories with matching grid map. Although this technique used a DL approach to learn the future movements of the pedestrian, it does not use a data-driven approach, which again can cause issues when experiencing unforeseen scenarios.

9.3. DL Approach

In [181], a data-driven approach using a deep network is proposed which uses the skeleton features of the pedestrian to estimate future intention. The evaluation provided results similar to [189] in terms of classification of whether the pedestrian will cross or stop when approaching the road, but the data-driven approach in [181] is a simpler method and requires less dense information, i.e., requiring monocular information rather than stereo and dense optical flow as in [178,189]. In other approaches, stereo cameras [191] and LiDAR [183] are used to predict pedestrian intentions based on their silhouette. Head and body orientation estimation have also been used for intention estimation for pedestrians in [190,192,193,216]. Fang et al [181] argue that it is unclear how these estimations provide accurate intention estimation or if they provide significant additional time for reactive manoeuvres. For example, before the collision, pedestrians typically look in the direction of the oncoming vehicle [196]. In [181] a pose estimation technique is employed and the orientation of the body and head of the pedestrian is considered as in [193]. It is suggested in [192], that head orientation is not particularly useful for pedestrians that are intending to stop or cross the road. Delays may be caused when predicting pedestrian intentions due to insufficient information on the posture and body movements [194], which also makes the data-driven approach more effective as it will use the information provided from the sensors in real time.

An approach using a vision-based technique to evaluate the pose of a pedestrian over several frames to establish the risk to the pedestrian with respect to the vehicle in [181]. The approach uses a CNN-based technique to detect and estimate the pose of pedestrian based on the work in [217]. It is also mentioned in [181] that high-level features, such as skeleton joints, can provide more information than low-levels, such as HOG and Histogram of Optical Flow (HOF) [218]. The overall design used a monocular camera for a pedestrian detector and 2D pedestrian pose estimation for determining the intentions of the pedestrian [219–221], whereas [220,221] used machine learning techniques that can also be applied using DL techniques. This technique is simpler to implement than some other pedestrian intention estimation techniques that require stereo cameras and optical flow to function. For future work, [181] suggests the application of the technique in situations where there are multiple pedestrians and with occlusion for evaluation. However, to achieve this application, a dataset would need to be produced to sufficiently evaluate the technique. In [222], a thermal sensor was used for more accurate results in low-light scenarios. The proposed method combined body features (e.g., standing, walking) with head orientation features. This method uses the distance between a pedestrian and curb (DPC), the lateral moving speed (LMS) and head orientation (HO) of the pedestrian. The approach provided the lowest error rate (22.03%) than any other combination of the features. This approach outperformed other prediction methods, such as a Markovian model [223] and a DBN approach [203].

RNNs have been used for sequence-based prediction in various applications, such as human gait analysis [224], handwriting imitation [225] and human interactions [226]. The RNN uses a feedback loop to capture temporal information by going into an internal state known as the hidden unit. RNNs allow for data to be fed back to the previous layers [183]. However, this type of RNN can become inaccurate in long-term predictions. Therefore, the RNN+LSTM architecture [207], which provides for a memory unit for the RNN in the LSTM units. The stored information values can change depending on the previous outputs and new inputs. These LSTM units consist of a cell state and four gate layers, with each gate consisting of an activation function that takes into account the current state and the previous state as an input. There can be multiple memory layers, depending on the application. The other three gates are the input, output and forget. The input gate selects the input that is sent to the memory, the output is based on the memory state and input and the forget gate selects the information that is to be discarded by the memory. A detailed discussion of the architecture of an RNN+LSTM architecture can be found in [202]. The proposed method in [202] was data-driven for long-term pedestrian intention prediction using a stacked LSTM architecture and evaluated on the Daimler pedestrian motion trajectory dataset with promising results for intention estimation.

10. Conclusions

Visible data is the typical type of data that is used for VRU detection and intent estimation. It is argued that visible data is not very robust on its own as its reliability diminishes in low-light conditions. It was suggested by many authors that the fusion of thermal data with visible data could improve the reliability and accuracy of detection, making a more robust overall system. Fusion techniques have provided positive results, but efforts will continue to find the ideal fusion technique. What is also lacking is the dataset for cyclists, both in visible and thermal datasets. On the other hand, detection techniques for pedestrians can be adapted for cyclists. Datasets for pedestrians and cyclists in a multispectral dataset can aid in improving the accuracy and speed of object detection techniques.

Detection is the preliminary phase for the purposes of intent estimation, enabling it to identify the pedestrian/cyclist in the surrounding environment. As detection has been a challenging problem in computer vision, there is a significant amount of literature on the topic. However, this remains a problem that is yet unsolved. DL aims to aid in overcoming this challenge. DL, largely CNN, has provided a more effective method of pedestrian and cyclist detection when compared to the traditional methods that were depended on hand-crafted descriptors for region proposal, feature extraction and classification. The ability to outperform the traditional approaches is partly due to the higher level of abstraction that is achieved by the deep network, which can be imitated through hand-crafted techniques. The techniques that we have mentioned need further attention so that they can operate in real time.

However, intent estimation techniques have not received the same attention. This is due to detection being the initial step to identifying the desired object. Once an object is successfully classified as the desired object, it can be tracked so that pose/orientation estimations can be examined. Using this information, an accurate intent estimation can be achieved. DL techniques are also being employed for intention estimation approaches as traditional models of path prediction and motion modelling are not sufficient. Traditional techniques are not as robust as the data-driven techniques based on RNN+LSTM methods. Data-driven can react to unseen situations, these reactions enable it to more effective and accurate in real time.

Author Contributions: S.A. is the main author of the current review article. All authors made substantial contributions to conception and design, participated in drafting the article or revising it critically for important intellectual content; and gave final approval of the version to be submitted and any revised version.

Funding: This research received no external funding.

Acknowledgments: We would like to acknowledge the support of James Spooner from Centre for Connected and Autonomous Automotive Research, Coventry University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dominguez-Sanchez, A.; Cazorla, M.; Orts-Escolano, S. Pedestrian Movement Direction Recognition Using Convolutional Neural Networks. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 3540–3548. [CrossRef]
2. Prabhakar, G.; Kailath, B.; Natarajan, S.; Kumar, R. Obstacle detection and classification using deep learning for tracking in high-speed autonomous driving. In Proceedings of the 2017 IEEE Region 10 Symposium (TENSYMP), Cochin, India, 14–16 July 2017; pp. 1–6. [CrossRef]
3. Tumas, P.; Jonkus, A.; Serackis, A. Acceleration of HOG based Pedestrian Detection in FIR Camera Video Stream. In Proceedings of the 2018 Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 26 April 2018; pp. 1–4.
4. Savasturk, D.; Froehlich, B.; Schneider, N.; Enzweiler, M.; Franke, U. A Comparison Study on Vehicle Detection in Far Infrared and Regular Images. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, ITSC, Las Palmas, Spain, 15–18 September 2015; pp. 1595–1600. [CrossRef]
5. Ren, J.; Chen, X.; Liu, J.; Sun, W.; Pang, J.; Yan, Q.; Tai, Y.W.; Xu, L. Accurate Single Stage Detector Using Recurrent Rolling Convolution. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
6. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
7. Szegedy, C.; Liu, W.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 1701–1708.
8. Hu, Q.; Wang, P.; Shen, C.; Van Den Hengel, A.; Porikli, F. Pushing the Limits of Deep CNNs for Pedestrian Detection. *IEEE Trans. Circ. Syst. Video Technol.* **2017**, *28*, 1358–1368. [CrossRef]
9. Tompson, J.; Jain, A.; Lecun, Y.; Bregler, C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *NIPS* **2014**, *2014*, 1799–1807.
10. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous Detection and Segmentation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 297–312.
11. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. An Evaluation of the State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 743–761. [CrossRef] [PubMed]
12. World Health Organisation. *Global Status Report on Road Safety 2015—Summary*; WHO: Geneva, Switzerland, 2015.
13. Toroyan, T.; Peden, M.M.; Iaych, K. Supporting a decade of action. *World Health Organisation* **2013**, *1*, 318. [CrossRef]
14. Bieshaar, M.; Reitberger, G.; Zernetsch, S.; Sick, B.; Fuchs, E.; Doll, K. Detecting Intentions of Vulnerable Road Users Based on Collective Intelligence. *arXiv* **2018**, arXiv:1809.03916.
15. European Commission. *2017 Road Safety Statistics: What Is behind the Figures*? European Commission: Brussels, Belgium, 2017.
16. Robineau, D. *Reported Road Casualties Great Britain, Annual Report 2017*; Department for Transport: London, UK, 2017.
17. Baek, J.; Hong, S.; Kim, J.; Kim, E. Efficient pedestrian detection at nighttime using a thermal camera. *Sensors* **2017**, *17*, 1850. [CrossRef]
18. European Road Safety Observatory. *Traffic Safety Basic Facts 2012*; Technical Report; European Road Safety Observatory: Athens, Greece: 2012.
19. Sun, W.; Zhu, S.; Ju, X.; Wang, D. Deep learning based pedestrian detection. In Proceedings of the Chinese Control And Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; pp. 1007–1011.
20. Gavrila, D.M.; Munder, S. Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle. *Int. J. Comput. Vis.* **2007**, *73*, 41–59. [CrossRef]
21. Gerónimo, D.; Sappa, A.D.; López, A.; Ponsa, D. Adaptive Image Sampling and Windows Classification for On—board Pedestrian Detection. In Proceedings of the International Conference on Computer Vision Systems, Bielefeld, Germany, 21–24 March 2007.

22. Shashua, A.; Gdalyahu, Y.; Hayun, G. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In Proceedings of the IEEE Intelligent Vehicles Symposium, Parma, Italy, 14–17 June 2004; pp. 1–6. [CrossRef]
23. Brunetti, A.; Buongiorno, D.; Trotta, G.F.; Bevilacqua, V. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing* **2018**, *300*, 17–33. [CrossRef]
24. Shaout, A.; Colella, D.M.; Awad, S.S. Advanced Driver Assistance Systems—Past, present and future. In Proceedings of the 2011 Seventh International Computer Engineering Conference (ICENCO'2011), Cairo, Egypt, 27–28 December 2011.
25. Li, X.; Li, L.; Flohr, F.; Wang, J.; Xiong, H.; Bernhard, M.; Pan, S.; Gavrila, D.M.; Li, K. A unified framework for concurrent pedestrian and cyclist detection. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 269–281. [CrossRef]
26. Wang, H.; Chen, Q.; Cai, W. Shape-based pedestrian/bicyclist detection via onboard stereo vision. In Proceedings of the Multiconference on "Computational Engineering in Systems Applications", Beijing, China, 4–6 Octpber 2006.
27. Noyce, D.A.; Dharmaraju, R.; Lehman, J.D. *An Evaluation of Technologies for Automated Detection and Classification of Pedestrians and Bicyclists*; Massachusetts Highway Department Report: Boston, MA, USA, 2002.
28. Solichin, A.; Harjoko, A.; Eko, A. A Survey of Pedestrian Detection in Video. *Int. J. Adv. Comput. Sci. Appl.* **2014**, *5*. [CrossRef]
29. Piccard, M. Background subtraction techniques: A review. *J. Hepatol.* **2004**. [CrossRef]
30. Benezeth, Y.; Jodoin, P.; Emile, B.; Laurent, H.; Rosenberger, C. Review and evaluation of commonly-implemented background subtraction algorithms. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4. [CrossRef]
31. Agarwal, S.; Terrail, J.O.D.; Jurie, F. Recent Advances in Object Detection in the Age of Deep Convolutional Neural Networks. *arXiv* **2018**, arXiv:1809.03193.
32. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
33. Chavez-Garcia, R.O.; Aycard, O. Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 525–534. [CrossRef]
34. Wang, X.; Han, T.X.; Yan, S. An HOG-LBP human detector with partial occlusion handling. In Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 32–39. [CrossRef]
35. Felzenszwalb, P.F.; Girshick, R.B.; Mcallester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
36. Roncancio, H.; Hernandes, A.C.; Becker, M. Vision-based system for pedestrian recognition using a tuned SVM classifier. In Proceedings of the Workshop on Engineering Applications, Bogota, Columbia, 2–4 May 2012. [CrossRef]
37. Yang, Y.; Liu, W.; Wang, Y.; Cai, Y. Research on the algorithm of pedestrian recognition in front of the vehicle based on SVM. In Proceedings of the 11th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, DCABES 2012, Guilin, China, 19–22 October 2012; pp. 396–400. [CrossRef]
38. Min, K.; Son, H.; Choe, Y.; Kim, Y.G. Real-time pedestrian detection based on A hierarchical two-stage Support Vector Machine. In Proceedings of the 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA), Melbourne, Australia, 19–21 June 2013; pp. 114–119. [CrossRef]
39. Neagoe, V.E.; Ciotec, A.D.; Bărar, A.P. A Concurrent Neural Network Approach to Pedestrian Detection in Thermal Imagery. In Proceedings of the 9th International Conference on Communications (COMM), Bucharest, Romania, 21–23 June 2012; pp. 133–136. [CrossRef]
40. Brazil, G.; Yin, X.; Liu, X. Illuminating Pedestrians via Simultaneous Detection and Segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4960–4969. [CrossRef]

41. Oren, M.; Papageorgiou, C.; Sinha, P.; Osuna, E.; Poggio, T. Pedestrian detection using wavelet templates. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; pp. 193–199. [CrossRef]
42. Viola, P.; Jones, M.J.; Snow, D. Detecting Pedestrians Using Patterns of Motion and Appearance. In Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 1, pp. 734–741. [CrossRef]
43. Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. *Ten Years of Pedestrian Detection, What Have We Learned?* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Basel, Switzerland, 2015; Volume 8926, pp. 613–627. [CrossRef]
44. Tomè, D.; Monti, F.; Baroffio, L.; Bondi, L.; Tagliasacchi, M.; Tubaro, S. Deep Convolutional Neural Networks for pedestrian detection. *Signal Process. Image Commun.* **2016**, *47*, 482–489. [CrossRef]
45. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
46. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8. [CrossRef]
47. Cho, H.; Rybski, P.E.; Bar-Hillel, A.; Zhang, W. Real-time pedestrian detection with deformable part models. In Proceedings of the 2012 IEEE Intelligent Vehicles Symposium, Alcala de Henares, Spain, 3–7 June 2012; pp. 1035–1042. [CrossRef]
48. Li, X.; Flohr, F.; Yang, Y.; Xiong, H.; Braun, M.; Pan, S.; Li, K.; Gavrila, D.M.; Flohr, F. A new benchmark for vision-based cyclist detection. In Proceedings of the IEEE Intelligent Vehicles Symposium, Gotenburg, Sweden, 19–22 June 2016; pp. 1028–1033.
49. Forsyth, D. Object detection with discriminatively trained part-based models. *Computer* **2014**, *47*, 6–7. [CrossRef]
50. Ghosh, S.; Amon, P.; Hutter, A.; Kaup, A. Reliable pedestrian detection using a deep neural network trained on pedestrian counts. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 685–689. [CrossRef]
51. Girshick, R.; Donahue, J.; Member, S.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef]
52. Cho, H.; Rybski, P.E.; Zhang, W. Vision-based bicyclist detection and tracking for intelligent vehicles. In Proceedings of the 2010 IEEE Intelligent Vehicles Symposium, San Diego, CA, USA, 21–24 June 2010; pp. 454–461. [CrossRef]
53. Tian, W.; Lauer, M. Fast and Robust Cyclist Detection for Monocular Camera Systems. In Proceedings of the International joint Conference on Computer Vision Imaging and Computer Graphics Theory and Applications (VISIGRAPP), Berlin, Germany, 11–14 March 2015.
54. Tong L.; Cao, X.; Yanwu X. An effective crossing cyclist detection on a moving vehicle. In Proceedings of the 2010 8th World Congress on Intelligent Control and Automation, Jinan, China, 7–9 July 2010; pp. 368–372. [CrossRef]
55. Mitter, C.S. Autonomous Car Ingredients: Safety, Surveillance and Infotainment, Part Three. *Sensors Magazine*, 7 September 2017.
56. *JdeRobot_DetectionSuite*; ROI_HOG; GitHub Inc.: San Francisco, CA, USA, 2018.
57. Hariyono, J.; Hoang, V.D.; Jo, K.H. Moving Object Localization Using Optical Flow for Pedestrian Detection from a Moving Vehicle. *Sci. World J.* **2014**, *2014*, 1–8. [CrossRef]
58. Mukherjee, R. Classification. 2018.
59. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*. [CrossRef] [PubMed]
60. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
61. Nam, W.; Dollár, P.; Han, J.H. Local Decorrelation For Improved Detection. *arXiv* **2014**, arXiv:1406.1134
62. Nguyen, D.T.; Li, W.; Ogunbona, P.O. Human detection from images and videos: A survey. *Pattern Recognit.* **2016**, *51*, 148–175. [CrossRef]

63. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708. [CrossRef]
64. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1106–1114. [CrossRef]
65. Ordóñez, F.J.; Roggen, D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115. [CrossRef]
66. Palaz, D.; Doss, M.M.; Collobert, R. Analysis of CNN-based Speech Recognition System using Raw Speech as Input. In Proceedings of the Conference of International Speech Communication Association (Interspeech), Dresden, Germang, 6–10 Septermber 2015; pp. 11–15.
67. Bo Yang, J.; Nhut Nguyen, M.; Phyo San, P.; Li Li, X.; Krishnaswamy, S. Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition. In Proceedings of the International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
68. Dollár, P.; Babenko, B.; Belongie, S.; Perona, P.; Tu, Z. Multiple Component Learning for Object Detection. In Proceedings of the ECCV, Marseille, France, 12–18 October 2008.
69. Ess, A.; Leibe, B.; Schindler, K.; Van Gool, L. A mobile vision system for robust multi-person tracking. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8. [CrossRef]
70. Lin, Z.; Davis, L.S. A Pose-Invariant Descriptor for Human Detection and Segmentation. In Proceedings of the Computer Vision—ECCV 2008, Marseille, France, 12–18 October 2008; pp. 423–436. [CrossRef]
71. Munder, S.; Schnorr, C.; Gavrila, D. Pedestrian Detection and Tracking Using a Mixture of View-Based Shape–Texture Models. *IEEE Trans. Intell. Transp. Syst.* **2008**, *9*, 333–343. [CrossRef]
72. Sabzmeydani, P.; Mori, G. Detecting Pedestrians by Learning Shapelet Features. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8. [CrossRef]
73. Seemann, E.; Fritz, M.; Schiele, B. Towards Robust Pedestrian Detection in Crowded Image Sequences. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8. [CrossRef]
74. Sharma, V.; Davis, J.W. Integrating Appearance and Motion Cues for Simultaneous Detection and Segmentation of Pedestrians. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8. [CrossRef]
75. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: The State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761. [CrossRef]
76. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:1207.0580.
77. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. *ACM* **2014**, 675–678. [CrossRef]
78. Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**. [CrossRef]
79. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the EEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
80. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; Lecun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Scottsdale, Arizona, 2–4 May 2013.
81. Toshev, A.; Christian Szegedy, G. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 1653–1660.
82. Sainath, T.N.; Mohamed, A.; Kingsbury, B.; Ramabhadran, B. Deep Convolutional Neural Networks for LVCSR. *Scand. J. Rheumatol.* **2015**, 39–48. [CrossRef]
83. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Proceedings of the IEEE conference on Advances in Neural Information Processing, Barcelona, Spain, 5–10 December 2016; pp. 379–387.

84. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
85. Song, H.; Choi, I.K.; Ko, M.S.; Bae, J.; Kwak, S.; Yoo, J. Vulnerable pedestrian detection and tracking using deep learning. In Proceedings of the 2018 International Conference on Electronics, Information, and Communication (ICEIC), Honolulu, HI, USA, 24–27 January 2018; pp. 1–2. [CrossRef]
86. Hou, Y.L.; Song, Y.; Hao, X.; Shen, Y.; Qian, M. Multispectral pedestrian detection based on deep convolutional neural networks. In Proceedings of the IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Xiamen, China, 22–25 October 2017.
87. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640.
88. Angelova, A.; Krizhevsky, A.; Vanhoucke, V. Pedestrian detection with a Large-Field-Of-View deep network. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 704–711. [CrossRef]
89. Tian, Y.; Luo, P.; Wang, X.; Tang, X. Deep learning strong parts for pedestrian detection. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1904–1912. [CrossRef]
90. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 354–370.
91. Wang, L.; Xu, L.; Yang, M.H. Pedestrian detection in crowded scenes via scale and occlusion analysis. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 1210–1214. [CrossRef]
92. Du, X.; El-Khamy, M.; Lee, J.; Davis, L. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, 24–31 March 2017; pp. 953–961. [CrossRef]
93. Xiang, Y.; Choi, W.; Lin, Y.; Savarese, S. Subcategory-aware Convolutional Neural Networks for Object Proposals and Detection. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017.
94. Yang, F.; Choi, W.; Lin, Y. Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2129–2137. [CrossRef]
95. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision(ICCV), Santiago, Chile, 7–13 December 2015.
96. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 346–361.
97. González, A.; Fang, Z.; Socarras, Y.; Serrat, J.; Vázquez, D.; Xu, J.; López, A.M. Pedestrian Detection at Day/Night Time with Visible and FIR Cameras: A Comparison. *Sensors* **2016**, *16*, 820. [CrossRef] [PubMed]
98. Erhan, D.; Szegedy, C.; Toshev, A.; Anguelov, D. Scalable Object Detection Using Deep Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2155–2162. [CrossRef]
99. Szegedy, C.; Reed, S.; Erhan, D.; Anguelov, D.; Ioffe, S. Scalable, High-Quality Object Detection. *arXiv* **2015**, arXiv:1412.1441.
100. Hosang, J.; Benenson, R.; Dollar, P.; Schiele, B. What Makes for Effective Detection Proposals? *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 814–830. [CrossRef]
101. Gerónimo, D.; López, A.M.; Sappa, A.D.; Graf, T. Survey of Pedestrian Detection for Advanced Driver Assistance Systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1239–1258. [CrossRef] [PubMed]
102. Enzweiler, M.; Gavrila, D.M. Monocular Pedestrian Detection: Survey and Experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2179–2195. [CrossRef] [PubMed]
103. Dolí, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**. [CrossRef]

104. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361. [CrossRef]
105. Leibe, B.; Seemann, E.; Schiele, B. Pedestrian Detection in Crowded Scenes. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 878–885. [CrossRef]
106. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So, I. Multispectral Pedestrian Detection: Benchmark Dataset and Baseline. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.
107. Bertozzi, M.; Broggi, A.; Felisa, M.; Vezzoni, G.; Del Rose, M. Low-level Pedestrian Detection by means of Visible and Far Infra-red Tetra-vision. In Proceedings of the IEEE Intelligent Vehicles Symposium, Tokyo, Japan, 13–15 June 2006; pp. 231–236. [CrossRef]
108. Scheunert, U.; Cramer, H.; Fardi, B.; Wanielik, G. Multi sensor based tracking of pedestrians: A survey of suitable movement models. In Proceedings of the IEEE Intelligent Vehicles Symposium, Parma, Italy, 14–17 June 2004; pp. 774–778. [CrossRef]
109. Yunus, K.R.; Mechkul, M.A. Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking. *Int. Res. J. Eng. Technol. (IRJET)* **2017**, *4*. [CrossRef]
110. Ziebinski, A.; Cupek, R.; Erdogan, H.; Waechter, S. A Survey of ADAS Technologies for the Future Perspective of Sensor Fusion. In Proceedings of the International Conference on Computational Collective Intelligence, Halkidiki, Greece, 28–30 September 2016; pp. 135–146. [CrossRef]
111. Hyun, E.; Jin, Y.S.; Lee, J.H.; Hyun, E.; Jin, Y.S.; Lee, J.H. A Pedestrian Detection Scheme Using a Coherent Phase Difference Method Based on 2D Range-Doppler FMCW Radar. *Sensors* **2016**, *16*, 124. [CrossRef]
112. Werling, M.; Thrun, S.; Levinson, J.; Askeland, J.; Becker, J.; Dolson, J.; Held, D.; Kammel, S.; Zico, J.K.; Langer, D.; et al. Towards Fully Autonomous Driving: Systems and Algorithms. In Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV), Baden, Germany, 5–9 June 2011. [CrossRef]
113. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. *Mach. Vis. Appl.* **2014**, *25*, 245–262. [CrossRef]
114. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation: Supplementary Material. *arXiv* **2014**, arXiv:1407.5736.
115. Eitel, A.; Springenberg, J.T.; Spinello, L.; Riedmiller, M.; Burgard, W. Multimodal deep learning for robust RGB-D object recognition. In Proceedings of the International Conference on Intelligent Robots and Systems, Hamburg, Germany, 28 September–2 October 2015; pp. 681–687. [CrossRef]
116. Xia, Y.; Xu, W.; Zhang, L.; Shi, X.; Mao, K. Integrating 3D structure into traffic scene understanding with RGB-D data. *Neurocomputing* **2015**, *151*, 700–709. [CrossRef]
117. Sun, H.; Wang, C.; Wang, B.; El-Sheimy, N. Pyramid binary pattern features for real-time pedestrian detection from infrared videos. *Neurocomputing* **2011**, *74*, 797–804. [CrossRef]
118. St-Laurent, L.; Maldague, X.; Prévost, D. Combination of colour and thermal sensors for enhanced object detection. In Proceedings of the FUSION 2007—2007 10th International Conference on Information Fusion, Quebec, QC, Canada, 9–12 July 2007. [CrossRef]
119. Socarrás, Y.; Ramos, S.; Vázquez, D.; López, A.M.; Gevers, T. Adapting Pedestrian Detection from Synthetic to Far Infrared Images. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
120. Miron, A.; Rogozan, A.; Ainouz, S.; Bensrhair, A.; Broggi, A. An Evaluation of the Pedestrian Classification in a Multi-Domain Multi-Modality Setup. *Sensors* **2015**, *15*, 13851–13873. [CrossRef]
121. Li, X.; Guo, R.; Chen, C. Robust Pedestrian Tracking and Recognition from FLIR Video: A Unified Approach via Sparse Coding. *Sensors* **2014**, *14*, 11245–11259. [CrossRef] [PubMed]
122. Besbes, B.; Rogozan, A.; Rus, A.M.; Bensrhair, A.; Broggi, A. Pedestrian detection in far-infrared daytime images using a hierarchical codebook of SURF. *Sensors* **2015**, *15*, 8570–8594. [CrossRef]
123. Lee, J.H.; Choi, J.S.; Jeon, E.S.; Kim, Y.G.; Le, T.T.; Shin, K.Y.; Lee, H.C.; Park, K.R. Robust pedestrian detection by combining visible and thermal infrared cameras. *Sensors* **2015**, *15*, 10580–10615. [CrossRef]
124. Senart, A.; Karpinski, M.; Wieckowski, M.; Cahill, V. Using Sensor Networks for Pedestrian Detection. In Proceedings of the 2008 5th IEEE Consumer Communications and Networking Conference, Las Vegas, NV, USA, 10–12 January 2008; pp. 697–701. [CrossRef]

125. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral Deep Neural Networks for Pedestrian Detection. In Proceedings of the British Machine Vision Conference, York, UK, 19–22 September 2016; pp. 1–13.
126. Kang, J.K.; Hong, H.G.; Park, K.R. Pedestrian detection based on adaptive selection of visible light or far-infrared light camera image by fuzzy inference system and convolutional neural network-based verification. *Sensors* **2017**, *17*, 1598. [CrossRef]
127. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully Convolutional Region Proposal Networks for Multispectral Person Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 243–250. [CrossRef]
128. Wagner, J.; Fischer, V.; Herman, M. Multispectral pedestrian detection using deep fusion convolutional neural networks. In Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 27–29 April 2016.
129. Luo, P.; Tian, Y.; Wang, X.; Tang, X. Switchable Deep Network for Pedestrian Detection. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 899–906. [CrossRef]
130. Tian, Y.; Luo, P.; Wang, X.; Tang, X. Pedestrian detection aided by deep learning semantic tasks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5079–5087. [CrossRef]
131. Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; Yan, S. Scale-aware Fast R-CNN for Pedestrian Detection. *IEEE Trans. Multimed.* **2017**, 1. [CrossRef]
132. Kruthiventi, S.S.S.; Sahay, P.; Biswal, R. Low-light pedestrian detection from RGB images using multi-modal knowledge distillation. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 4207–4211. [CrossRef]
133. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral Deep Neural Networks for Pedestrian Detection. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 27–29 April 2016.
134. Choi, H.; Kim, S.; Park, K.; Sohn, K. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In Proceedings of the International Conference on Pattern Recognition, Cancun, Mexico, 4–8 December 2017; pp. 621–626. [CrossRef]
135. Park, K.; Kim, S.; Sohn, K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognit.* **2018**, *80*, 143–155. [CrossRef]
136. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
137. RANCHIN, T.; WALD, L. The wavelet transform for the analysis of remotely sensed images. *Int. J. Remote Sens.* **1993**, *14*, 615–619. [CrossRef]
138. Gao, H.; Zou, B. Algorithms of image fusion based on wavelet transform. In Proceedings of the 2012 International Conference on Image Analysis and Signal Processing, Hangzhou, China, 9–11 November 2012; pp. 1–4. [CrossRef]
139. Candès, E.; Demanet, L.; Donoho, D.; Ying, L. Fast Discrete Curvelet Transforms. *Multiscale Model. Simul.* **2006**, *5*, 861–899. [CrossRef]
140. Burt, P.; Adelson, E. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. Commun.* **1983**, *31*, 532–540. [CrossRef]
141. Torresan, H.; Turgeon, B.; Ibarra-Castanedo, C.; Hebert, P.; Maldague, X.P. Advanced surveillance systems: Combining video and thermal imagery for pedestrian detection. In Proceedings of the SPIE, Thermosense XXVI, Orlando, FL, USA, 12 April 2004; Burleigh, D.D., Cramer, K.E., Peacock, G.R., Eds.; International Society for Optics and Photonics; Volume 5405; pp. 506–515. [CrossRef]
142. Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J.S.; An, K.; Kweon, I.S. KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 934–948. [CrossRef]
143. Deng, J.D.J.; Dong, W.D.W.; Socher, R.; Li, L.J.L.L.J.; Li, K.L.K.; Fei-Fei, L.F.F.L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2–9. [CrossRef]
144. Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; Winnemoeller, H. Recognizing Image Style. *Comput. Vis. Pattern Recognit.* **2014**. [CrossRef]

145. Dollar, P.; Tu, Z.; Tao, H.; Belongie, S. Feature Mining for Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8. [CrossRef]
146. Schwartz, W.R.; Kembhavi, A.; Harwood, D.; Davis, L.S. Human detection using partial least squares analysis. In Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 November 2009; pp. 24–31. [CrossRef]
147. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
148. Dollár, P.; Belongie, S.; Perona, P. The Fastest Pedestrian Detector in the West. In Proceedings of the British Machine Vision Conference, Aberystwyth, UK, 30 August–2 September 2010.
149. Wojek, C.; Walk, S.; Schiele, B. Multi-cue onboard pedestrian detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 794–801. [CrossRef]
150. Dollar, P.; Tu, Z.; Perona, P.; Belongie, S. Integral Channel Features. In Proceedings of the British Machine Vision Conference 2009, London, UK, 7–10 September 2009; pp. 1–91. [CrossRef]
151. Walk, S.; Majer, N.; Schindler, K.; Schiele, B. New features and insights for pedestrian detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1030–1037. [CrossRef]
152. Wojek, C.; Schiele, B. A Performance Evaluation of Single and Multi-feature People Detection. *Pattern Recognit.* **2008**, 82–91. [CrossRef]
153. Maji, S.; Berg, A.C.; Malik, J. Classification using intersection kernel support vector machines is efficient. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8. [CrossRef]
154. Ess, A.; Leibe, B.; Van Gool, L. Depth and Appearance for Mobile Scene Analysis. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8. [CrossRef]
155. Ouyang, W.; Wang, X. Joint Deep Learning for Pedestrian Detection. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2056–2063. [CrossRef]
156. Sermanet, P.; Kavukcuoglu, K.; Chintala, S.; Lecun, Y. Pedestrian Detection with Unsupervised Multi-stage Feature Learning. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3626–3633. [CrossRef]
157. Woonhyun N.; Bohyung H.; Joon Hee H. Improving object localization using macrofeature layout selection. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1801–1808. [CrossRef]
158. Costea, A.D.; Nedevschi, S. Word Channel Based Multiscale Pedestrian Detection without Image Resizing and Using Only One Classifier. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 899–906; pp. 2393–2400. [CrossRef]
159. Ţoca, C.; Ciuc, M.; Pătraşcu, C. Normalized Autobinomial Markov Channels For Pedestrian Detection. In Procedings of the British Machine Vision Conference 2015, Swansea, UK, 7–10 September 2015; pp. 1–175. [CrossRef]
160. Marin, J.; Vazquez, D.; Lopez, A.M.; Amores, J.; Leibe, B. Random Forests of Local Experts for Pedestrian Detection. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2592–2599. [CrossRef]
161. Yang, Y.; Wang, Z.; Wu, F. Exploring Prior Knowledge for Pedestrian Detection. In Proceedings of the British Machine Vision Conference 2015, Swansea, UK, 7–10 September 2015; pp. 1–176. [CrossRef]
162. Mathias, M.; Benenson, R.; Timofte, R.; Gool, L.V. Handling Occlusions with Franken-Classifiers. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1505–1512. [CrossRef]
163. Zhang, S.; Bauckhage, C.; Cremers, A.B. Informed Haar-like Features Improve Pedestrian Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014, pp. 947–954.

164. Benenson, R.; Mathias, M.; Tuytelaars, T.; Van Gool, L. Seeking the Strongest Rigid Detector. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3666–3673. [CrossRef]
165. Lim, J.J.; Zitnick, C.L.; Dollar, P. Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3158–3165. [CrossRef]
166. Paisitkriangkrai, S.; Shen, C.; Hengel, A.v.d. Pedestrian Detection with Spatially Pooled Features and Structured Ensemble Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1243–1257. [CrossRef]
167. Zhang, L.; Lin, L.; Liang, X.; He, K. Is Faster R-CNN Doing Well for Pedestrian Detection? In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 443–457.
168. Ouyang, W.; Member, S.; Zhou, H.; Li, H. Jointly Learning Deep Features , Deformable Parts , Occlusion and Classification for Pedestrian Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1874–1887. [CrossRef] [PubMed]
169. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3339–3348. [CrossRef]
170. Du, X.; El-Khamy, M.; Morariu, V.I.; Lee, J.; Davis, L.S. Fused Deep Neural Networks for Efficient Pedestrian Detection. *arXiv* **2018**, arXiv:1805.08688.
171. Song, T.; Sun, L.; Xie, D.; Sun, H.; Pu, S. Small-scale Pedestrian Detection Based on Somatic Topology Localization and Temporal Feature Aggregation. *arXiv* **2018**, arXiv:1807.01438.
172. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. *Caltech Pedestrian Detection Benchmark*; IEEE Conference on Computer Vision and Pattern Recognition: Florida, 2012.
173. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: A benchmark. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 304–311. [CrossRef]
174. Ess, A.; Leibe, B.; Schindler, K.; van Gool, L. Robust multiperson tracking from a mobile platform. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 1831–1846. [CrossRef] [PubMed]
175. Leibe, E. Multi-Cue Onboard Pedestrian Detection. In Proceedings of the CVPR, Miami, FL, USA, 20–25 June 2009; pp. 794–801.
176. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
177. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016 2016; pp. 3213–3223. [CrossRef]
178. Schneider, N.; Gavrila, D.M. Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study. In Proceedings of the Conference on Pattern Recognition. Springer, Berlin, Heidelberg, Saarbrücken, Germany, 3–6 September 2013; pp. 174–183. [CrossRef]
179. Gauen, K.; Dailey, R.; Laiman, J.; Zi, Y.; Asokan, N.; Lu, Y.H.; Thiruvathukal, G.K.; Shyu, M.L.; Chen, S.C. Comparison of Visual Datasets for Machine Learning. In Proceedings of the 2017 IEEE International Conference on Information Reuse and Integration (IRI), San Diego, CA, USA, 4–6 August 2017; pp. 346–355. [CrossRef]
180. FLIR. *FLIR Releases Starter Thermal Imaging Dataset for Machine Learning Advanced Driver Assistance Development*; FLIR Systems: Wilsonville, OR, USA, 2018.
181. Fang, Z.; Vázquez, D.; López, A.; Fang, Z.; Vázquez, D.; López, A.M. On-Board Detection of Pedestrian Intentions. *Sensors* **2017**, *17*, 2193. [CrossRef]
182. Kohler, S.; Goldhammer, M.; Bauer, S.; Doll, K.; Brunsmann, U.; Dietmayer, K. Early detection of the Pedestrian's intention to cross the street. In Proceedings of the 2012 15th International IEEE Conference on Intelligent Transportation Systems, Anchorage, AK, USA, 16–19 September 2012; pp. 1759–1764. [CrossRef]
183. Volz, B.; Behrendt, K.; Mielenz, H.; Gilitschenski, I.; Siegwart, R.; Nieto, J. A data-driven approach for pedestrian intention estimation. In Proceedings of the International Conference on Intelligent Transportation Systems. IEEE, Rio de Janeiro, Brazil, 1–4 November 2016; pp. 2607–2612. [CrossRef]

184. López, A.M.; Imiya, A.; Pajdla, T.; Álvarez, J.M. Computer Vision for MAVs. In *Computer Vision in Vehicle Technology: Land, Sea and Air*; Wiley: New Jersey, NJ, USA, 2017; Chapter 3, pp. 24–54. [CrossRef]
185. Braeuchle, C.; Ruenz, J.; Flehmig, F.; Rosenstiel, W.; Kropf, T. Situation analysis and decision making for active pedestrian protection using Bayesian networks. In Proceedings of the 6. Tagung Fahrerassistenz München, München, Germany, 28–29 November 2013.
186. Hoermann, S.; Bach, M.; Dietmayer, K. Dynamic Occupancy Grid Prediction for Urban Autonomous Driving: A Deep Learning Approach with Fully Automatic Labeling. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 2056–2063. [CrossRef]
187. Raza, M.; Chen, Z.; Rehman, S.U.; Wang, P.; Bao, P. Appearance based pedestrians' head pose and body orientation estimation using deep learning. *Neurocomputing* **2018**, *272*, 647–659. [CrossRef]
188. Koehler, S.; Goldhammer, M.; Bauer, S.; Zecha, S.; Doll, K.; Brunsmann, U.; Dietmayer, K. Stationary Detection of the Pedestrian's Intention at Intersections. *IEEE Intell. Transp. Syst. Mag.* **2013**, *5*, 87–99. [CrossRef]
189. Keller, C.G.; Gavrila, D. Will the Pedestrian Cross? A Study on Pedestrian Path Prediction. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 494–506. [CrossRef]
190. Rehder, E.; Kloeden, H.; Stiller, C. Head detection and orientation estimation for pedestrian safety. In Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; pp. 2292–2297. [CrossRef]
191. Kohler, S.; Goldhammer, M.; Zindler, K.; Doll, K.; Dietmeyer, K. Stereo-Vision-Based Pedestrian's Intention Detection in a Moving Vehicle. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas, Spain, 15–18 September 2015; pp. 2317–2322. [CrossRef]
192. Schulz, A.T.; Stiefelhagen, R. Pedestrian intention recognition using Latent-dynamic Conditional Random Fields. In Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (IV), Seoul, Korea, 28 June–1 July 2015; pp. 622–627. [CrossRef]
193. Flohr, F.; Dumitru-Guzu, M.; Kooij, J.F.P.; Gavrila, D.M. A Probabilistic Framework for Joint Pedestrian Head and Body Orientation Estimation. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 1872–1882. [CrossRef]
194. Schneemann, F.; Heinemann, P. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Daejeon, Korea 9–14 October 2016; pp. 2243–2248. [CrossRef]
195. Kwak, J.Y.; Lee, E.J.; Ko, B.; Jeong, M. Pedestrian's Intention Prediction Based on Fuzzy Finite Automata and Spatial-temporal Features. In Proceedings of the International Symposium on Electronic Imaging—Video Surveillance and Transportation Imaging Applications, San Francisco, CA, USA, 14–18 February 2016. [CrossRef]
196. Rasouli, A.; Kotseruba, I.; Tsotsos, J.K. Agreeing to cross: How drivers and pedestrians communicate. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 264–269. [CrossRef]
197. Quintero, R.; Parra, I.; Lorenzo, J.; Fernandez-Llorca, D.; Sotelo, M.A. Pedestrian intention recognition by means of a Hidden Markov Model and body language. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 1–7. [CrossRef]
198. Furuhashi, R.; Yamada, K. Estimation of street crossing intention from a pedestrian's posture on a sidewalk using multiple image frames. In Proceedings of the First Asian Conference on Pattern Recognition, Beijing, China, 28 November 2011; pp. 17–21. [CrossRef]
199. Quintero, R.; Parra, I.; Llorca, D.F.; Sotelo, M.A. Pedestrian Intention and Pose Prediction through Dynamical Models and Behaviour Classification. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas, Spain, 15–18 September 2015; pp. 83–88. [CrossRef]
200. Rehder, E.; Kloeden, H. Goal-Directed Pedestrian Prediction. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, 7–13 December 2015; pp. 139–147. [CrossRef]

201. Keller, C.G.; Hermes, C.; Gavrila, D.M. *Will the Pedestrian Cross? Probabilistic Path Prediction Based on Learned Motion Features*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Basel, Switzerland, 2011; Volume 6835 LNCS, pp. 386–395. [CrossRef]
202. Saleh, K.; Hossny, M.; Nahavandi, S. Intent prediction of vulnerable road users from motion trajectories using stacked LSTM network. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 327–332. [CrossRef]
203. Kooij, J.F.P.; Schneider, N.; Flohr, F.; Gavrila, D.M. Context-Based Pedestrian Path Prediction. In Proceedings of the European Conference on ComputerVision (ECCV), Zurich, Switzerland, 6–12 September 2014, pp. 618–633. [CrossRef]
204. Liu, Q.; He, Z. PTB-TIR: A Thermal Infrared Pedestrian Tracking Benchmark. *arXiv* **2018**, arXiv:1801.05944.
205. Ma, Y.; Wu, X.; Yu, G.; Xu, Y.; Wang, Y.; Toro, F.G. Pedestrian Detection and Tracking from Low-Resolution Unmanned Aerial Vehicle Thermal Imagery. *Sensors* **2016**, *16*, 446. [CrossRef] [PubMed]
206. Volz, B.; Mielenz, H.; Agamennoni, G.; Siegwart, R. Feature Relevance Estimation for Learning Pedestrian Behavior at Crosswalks. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas, Spain, 15–18 September 2015; pp. 854–860. [CrossRef]
207. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
208. Ellis, D.; Sommerlade, E.; Reid, I. Modelling pedestrian trajectory patterns with Gaussian processes. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, Kyoto, Japan, 27 September–4 Octover 2009; pp. 1229–1234. [CrossRef]
209. Schmidt, S.; Färber, B. Pedestrians at the kerb–Recognising the action intentions of humans. *Transp. Res. Part F Traffic Psychol. Behav.* **2009**, *12*, 300–310. [CrossRef]
210. Quintero, R.; Almeida, J.; Llorca, D.F.; Sotelo, M.A. Pedestrian path prediction using body language traits. In Proceedings of the 2014 IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, USA, 8–11 June 2014; pp. 317–323. [CrossRef]
211. Fugger, T.F.; Randles, B.C.; Stein, A.C.; Whiting, W.C.; Gallagher, B. Analysis of Pedestrian Gait and Perception-Reaction at Signal-Controlled Crosswalk Intersections. *Transp. Res. Rec. J. Transp. Res. Board* **2000**, *1705*, 20–25. [CrossRef]
212. Goldhammer, M.; Hubert, A.; Koehler, S.; Zindler, K.; Brunsmann, U.; Doll, K.; Sick, B. Analysis on termination of pedestrians' gait at urban intersections. In Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; pp. 1758–1763. [CrossRef]
213. Ziebart, B.; Ratliff, N.; Gallagher, G. Planning-based prediction for pedestrians. In Proceedings of the EEE/RSJ International Conference on Intelligent Robots and Systems IROS, St. Louis, MO, USA, 10–15 October 2009; pp. 3931–3936.
214. Kitani, K.M.; Ziebart, B.D.; Bagnell, J.A.; Hebert, M. Activity Forecasting. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 201–214.
215. Zyner, A.; Worrall, S.; Ward, J.; Nebot, E. Long short term memory for driver intent prediction. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 1484–1489. [CrossRef]
216. Huang, Y.; Cui, J.; Davoine, F.; Zhao, H.; Zha, H. Head pose based intention prediction using Discrete Dynamic Bayesian Network. In Proceedings of the Seventh International Conference on Distributed Smart Cameras (ICDSC), Palm Springs, CA, USA, 29 October–1 November 2013; pp. 1–6. [CrossRef]
217. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
218. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards Understanding Action Recognition. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3192–3199. [CrossRef]
219. Enzweiler, M.; Gavrila, D.M. Integrated Pedestrian Classification and Orientation Estimation. In Proceedings of the CVPR, San Francisco, CA, USA, 13–18 June 2010; pp. 982–989.

220. Gandhi, T.; Trivedi, M.M. Image based estimation of pedestrian orientation for improving path prediction. In Proceedings of the 2008 IEEE Intelligent Vehicles Symposium, Eindhoven, The Netherlands, 4–6 June 2008; pp. 506–511. [CrossRef]
221. Mogelmose, A.; Trivedi, M.M.; Moeslund, T.B. Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations. In Proceedings of the 2015 IEEE Intelligent Vehicles Symposium (IV), Seoul, Korea, 28 June–1 July 2015; pp. 330–335. [CrossRef]
222. Kwak, J.Y.; Ko, B.C.; Nam, J.Y. Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime. *Infrared Phys. Technol.* **2017**, *81*, 41–51. [CrossRef]
223. Wakim, C.; Capperon, S.; Oksman, J. A Markovian model of pedestrian behavior. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, The Hague, The Netherlands, 10–13 October 2004; pp. 4028–4033. [CrossRef]
224. Fragkiadaki, K.; Levine, S.; Felsen, P.; Malik, J. Recurrent Network Models for Human Dynamics. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4346–4354. [CrossRef]
225. Graves, A. Generating Sequences With Recurrent Neural Networks. *arXiv* **2013**, arXiv:1308.0850.
226. Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; Savarese, S. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 26 June–1 July 2016; pp. 961–971.