Computer vision and deep learning techniques for pedestrian detection and tracking: A survey

Antonio Brunetti a, Domenico Buongiorno a, Gianpaolo Francesco Trotta b, Vitoantonio Bevilacqua a,∗

a Department of Electrical and Information Engineering (DEI), Polytechnic University of Bari, Italy
b Department of Mechanics, Mathematics and Management (DMMM), Polytechnic University of Bari, Italy

A R T I C L E   I N F O

Article history:
Received 15 September 2017
Revised 10 December 2017
Accepted 8 January 2018
Available online 10 March 2018

Keywords:
Pedestrian detection
Human tracking
Deep learning
Convolutional neural network
Machine learning
Artificial neural network
Features extraction

A B S T R A C T

Pedestrian detection and tracking have become an important field in the computer vision research area. This growing interest, started in the last decades, might be explained by the multitude of potential applications that could use the results of this research field, e.g. robotics, entertainment, surveillance, care for the elderly and disabled, and content-based indexing. In this survey paper, vision-based pedestrian detection systems are analysed based on their field of application, acquisition technology, computer vision techniques and classification strategies. Three main application fields have been individuated and discussed: video surveillance, human-machine interaction and analysis. Due to the large variety of acquisition technologies, this paper discusses both the differences between 2D and 3D vision systems, and indoor and outdoor systems. The authors reserved a dedicated section for the analysis of the Deep Learning methodologies, including the Convolutional Neural Networks in pedestrian detection and tracking, considering their recent exploding adoption for such a kind systems. Finally, focusing on the classification point of view, different Machine Learning techniques have been analysed, basing the discussion on the classification performances on different benchmark datasets. The reported results highlight the importance of testing pedestrian detection systems on different datasets to evaluate the robustness of the computed groups of features used as input to classifiers.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

The growing interest in autonomous cars demonstrated by the huge investments made by the biggest automotive and IT companies [1], as well as the development of machines and applications able to interact with persons [2–13], is playing an important role in the improvement of the techniques for vision-based pedestrian tracking. In fact, autonomous machines able to act in not-controlled environments represent an high risk for any person who may be in their range of action.

In 2015, in the United States, more than 5000 pedestrians were killed due to traffic crashes [14]: one pedestrian dies every 1.6 hours due to car accident. Additionally, in the same year, almost 130000 pedestrians were treated in emergency departments for non-fatal crash-related injuries. Pedestrians are 1.5 times more likely than passenger vehicle occupants to be killed in a car crash on each trip [14–17]. The statistics reported in [14] state alarming numbers for EU too, even though the general trend of pedestrians’ deaths is reducing thanks to the introduction of driving supports, such as auto-breaking system. For these reasons, in the last decades, people detection and tracking has become an important research area in computer vision.

From 1990 to 2016, scientific community has shown an ever-growing interest in human detection and tracking. As reported in Fig. 1, more than 5000 publications in this topic have been published and indexed in Web of Science, ranging from human detection to pedestrian tracking using 2D and 3D vision systems, or considering indoor and outdoor environments.

Some other surveys regarding pedestrian detection have been presented in the literature so far. In [18] and [19] the authors focused the topic of the survey on a taxonomy of system functionalities considering the structure of the motion capture system and the different information to be processed.

Solichin et al. have focused the work on the steps needed in the process of pedestrian detection, including input devices, datasets and methods for detection and, finally, on some open issues related to pedestrian detection [20].

Zhou and Hu have written a survey on the human detection and tracking systems from a clinical and diagnostic point of view,
highlighting the differences between visual tracking (marker-based or marker-less) and non-visual tracking using magnetic sensors, inertial sensors and others [21].

In [22], the authors have presented a survey concerning monocular pedestrian detection systems focusing on the methodologies for the selection of Regions Of Interest (ROIs), classification methods and tracking.

In [23] and [24], the authors have discussed two surveys focused on pedestrian detection and tracking systems related to the Pedestrian Protection Systems (PPSs). Specifically, while in the first survey the authors consider and review general pedestrian detectors, in the latter the authors focus only on systems dedicated to PPSs.

Dollar and his colleagues [25] have focused on the main methods for pedestrian detection in monocular images performing an accurate ranking on benchmark datasets, while in [26] the authors have collected and reviewed some works, marginally introducing deep architectures.

The above-mentioned surveys report the state-of-the-art about pedestrian detection and tracking systems in terms of acquisition technologies, e.g. 2D and 3D configurations, and processing methodologies; however, recent adoption of Deep Learning (DL) methodologies and, in particular, Convolutional Neural Networks (CNNs) for pedestrian detection and tracking deserves a dedicated state-of-the-art survey.

Generally, the process of vision-based pedestrian detection can be considered constituted by three fundamental steps, as depicted in Fig. 2: (i) Image Acquisition, (ii) Feature Extraction and (iii) Classification. As will be discussed in the following sections, the introduction of DL architectures, or deep structures inspired to the human visual cortex, in the context of object recognition, allowed the removal of the feature extraction step (Fig. 3), preserving the other ones.

As pointed out by the reported figures, the two approaches differ for the removed step only. However, the extraction of features is not completely removed from the workflow, but it is an automatic procedure performed by the deep classifier which is generally constituted by several processing layers that, taking images as input, compute features at different layers of abstraction [27–30]. In this way, the design of a such a kind classifiers is considerably simplified since the design of procedures for the extraction of the so called “hand-crafted features”, able to perform an accurate classification, is the most difficult step. Nevertheless, the incorporation of feature extraction in the classification process, allowing a faster run-time execution, lead to a longer training time respect to the hand-crafted features based approach [31].

Video tracking is a complex process which allows to locate and follow single or multiple objects over time using several sensors. Due to the need of a remarkable improvement in both acquisition and processing systems, a lot of works dealing with tracking could be found in literature. In fact, each moving object in the world could be potentially tracked regardless the tracking system. For example, complex systems based on radar or GPS are widely studied and used currently in different contexts, e.g. aviation industry or ground movements tracking [32–36].

Among the variety of traceable objects, human tracking is the most interesting since the processes of human detection and segmentation in images and videos are difficult due to the large variety of conditions and variables to take into account for this task, besides the well-known problems related to images segmentation, such as noise [37–41].

The automatic tracking of humans in video has always been an interesting research topic, as it is a cross-domain research area with infinite applications in different fields. In fact, the potential applications of human motion capture led to the development of systems in several domains, such as surveillance, control, and analysis. In addition, there are some recent research fields where the
automatic tracking of humans in video sequences is rising up, such as human-computer interaction and augmented reality [8,36,42–46].

Regardless of the kind of object to be tracked, the identification of Regions Of Interest (ROIs) is the first and most important step in the most of computer vision applications, including object tracking. This step requires the application of some image processing techniques in order to make easier the identification and selection of ROIs; the difficulty of this approach mostly depends on both the acquisition system (e.g., camera resolution, field of view and technology) and the environmental conditions (e.g., lighting conditions). Although it seems a trivial process, in some approaches the previous sequence of steps could be sufficient to track one or more objects into a video sequence under certain conditions [47–49].

In more complex applications, some features need to be extracted in order to describe the identified regions. The extracted features, whose kind is related to the acquired signal, are then used as input in the subsequent step for the discrimination of the identified objects; finally a tracker is necessary to follow the considered object (or class of objects) during the video flow [3,8,12,13,18,19,21–23,25,26,32,36,43,44,50–99].

To simplify and strengthen the step of ROIs identification, and consequently the overall tracking system, some authors introduced active and passive markers (or optical references) to be applied to the object to track. In literature, several kinds of marker could be found; their nature is strictly related to the technology of the acquisition system, allowing an accurate tracking in several conditions [9,90,100–109].

In recent years, the spread of innovative techniques based on Deep Learning has prompted many research groups to apply these techniques for the segmentation of objects in images and tracking in videos with different aims [110–119]. This kind of approach seems to be very interesting and powerful since the steps needed for the features extraction from the segmented ROIs is overcome thanks to DL architectures that make use of deep classifiers, such as Convolutional Neural Networks (CNNs).

In the sections that follow, we present the application fields of pedestrian detection and tracking systems, first. We then describe the different configurations of the vision systems in the Section 3. Subsequently, we present the different methods for video processing and features extraction in the Section 4 focusing on pedestrian subjects in the Section 4.1. Then, we introduce the approaches pedestrian classification using Machine Learning and Deep Learning techniques in the Section 5, while we present a final discussion in the Section 6. Finally, we present conclusions.

2. Applications

The growing interest for vision-based tracking systems can be explained by multiple factors. To the authors’ opinion, the most important factor is the advancement of the related fields that make use of the tracking techniques. In addition, recent researches with background in Artificial Intelligence (AI), Augmented Reality (AR) and medical imaging, as well as the diffusion of low cost video acquisition systems and more powerful processing devices, have contributed to the diffusion of researches in tracking systems.

The three major application areas individuated by the authors are: surveillance, human-machine interaction, and analysis.

2.1. Video surveillance

Surveillance applications based on human tracking are the most diffused in literature. The main goal is the detection of one or more people in the scene for tracking their movements in video flow over time. For example, several systems are able to monitor parking lots, airports or crowded places (Fig. 4).

The main differences that characterize the works found in literature about video surveillance applications consist in the acquisition systems (e.g., colour-space and resolution), the number of potentially traceable subjects (e.g., mono or multi target), and object categorization [51,61,65,77–79,91]. Among the video surveillance applications, the most afforded research topic is focused on pedestrian tracking (see Section 4.1 for more details) [94,97–99,116].

2.2. Human-machine interaction

The human-machine interaction area relates to tasks where the captured human motion is used to provide controlling functionalities for remote controlling and designing virtual game interfaces, virtual environments and animations.

Moreover, tracking systems have been also applied in the entertainment industry where the control of personalized graphic models is making the productions/products more realistic.

In recent years, the interest in using Unmanned Aerial Vehicles (UAVs) to accomplish a series of tasks that can be uncomfortable or dangerous to be performed by human beings has been subject to a constantly increase (Fig. 5). This can be mostly explained by the higher possibility to purchase a cheaper drone, especially for game and sport purposes. This has pushed the scientific community to investigate the capabilities of UAVs in video-based tracking applications [7,9,12,13,90,91,54,96–99,101–103,105,120–126].

Besides the UAV control, a lot of research and investments have been done by big corporations to support the research in self-driving cars [1,3,92].

2.3. Analysis

The analysis of captured motion data may be used in different clinical studies, e.g. to diagnose orthopaedic diseases, to help athletes in understanding and improving their performance, to restore patients’ functional capability in stroke rehabilitation or to prevent fall accidents [21,44,95,100,106–108]. In this kind of applications, the patients’ activities need to be continuously monitored, and subsequently corrected during motor-rehabilitation sessions [127–129].

Furthermore, these types of applications are used to answer questions about what people are doing and where and when they act. To achieve this goal, the algorithms implemented in these applications build people’s appearance patterns and trace people with relative identity (who) through occlusion events in the imagery. So they are used to increase awareness of security issue by performing analysis of actions, activities and behaviors both for crowds and individuals; for example, such systems are used for queue and shopping behavior analysis, detection of abnormal activities, and person identification [18,19,51–53,55,62,69,74,90,104,130].

3. Vision systems for pedestrian detection

The systems used to capture human motion consist of subsystems for sensing and processing, respectively. The operational complexity of these subsystems is typically related, i.e. the more complex the previous step is, the simpler the following one will be and vice versa. This trade-off between the complexities also relates to the use of active versus passive sensing.

Active sensing operates by placing devices on the subject and in the surroundings which transmit or receive generated signals. Active sensing allows for simpler processing and is widely used when the applications perform in well-controlled environments; for example, the most of applications in the analysis and control areas make use of active sensing (Fig. 6).

Passive sensing is based on “natural” signal sources, e.g. visual light or other electromagnetic wavelengths, and generally requires
no wearable devices. An exception can be make if markers are attached to the subject for an easier motion capture process. Visual markers are not as intrusive as the devices used in active sensing where passive sensing is mainly used in surveillance and some control applications where mounting any kind of active device on the subject it is not allowed.

Computer vision applications based on the passive sensing approach have challenged active sensing within all the considered application areas. Even though the use of markers could be a good compromise between passive and active sensing, the application of any kind of passive or active marker in real situations for tracking objects in uncontrolled or random environments is generally inconvenient or often impossible. For these reasons, systems able to detect and track objects considering only the acquired images from the vision system are needed.

Considering passive sensing, currently the great majority of the algorithms that accomplish similar tasks relies on colour information or on the use of external devices that track the target position [121,124–126].

Regardless of the use of active or passive sensors, vision systems for pedestrian detection may be differentiated considering the video acquisition technology (2D vs 3D), or the environmental conditions (Indoor vs Outdoor). These two topics will be afforded in the following two sections.

3.1. 2D vs 3D

Video acquisition technology is one of the fundamental aspects that concern with pedestrian, and generally, object detection and tracking. In literature, most of the works dealing with
pedestrian detection use a 2D acquisition system and Machine Learning (ML) techniques to perform a large variety of tasks [12,13,21–23,25,26,50,51,54,56,59–64,67–69,71,72,74–81,84–91,93,94,96–99,131–134].

In the most of cases, a 2D video is sufficient for pedestrian detection since videos contain extremely valuable information that can be extracted after an appropriate processing, i.e. the 2D coordinates of a detected person. Moreover, as will be discussed in the following section (Section 4), motion information, which is generally extracted using a 3D vision system, could be obtained following a bi-dimensional approach [3,18,19,21,24,44,52,53,55,65,76,92,95,135,136].

Conversely, in applications where the motion of a person should be acquired with high levels of accuracy, like pedestrian protection systems for autonomous vehicles or clinical and diagnostic environments for gait analysis, 3D motion capture systems are needed since 2D acquisition could lead to an excessive loss of information. 3D systems are able to generate a pedestrian coordinates representation in the 3D space (x, y, z planes of motion), often used to generate a representational model in virtual environments. This kind of systems generally differ for the sensing device that affect the overall cost of the application; in particular, 3D systems may use RGB-D cameras, which consist in the combination of a classic RGB camera and a Depth camera based on infra-red (IR) light acquisition [137,138]; stereo camera which is a type of camera constituted by two or more lenses with a separate image sensor or film frame for each lens to simulate human binocular vision, and therefore gives the ability to capture three dimensional images, by using stereo photograph. Multi-camera systems allows to capture the scene from several points of view and needs, in the latter case, calibration and registration phases and optical markers to obtain high levels of accuracy during people tracking.

Typically, 3D optical systems are more expensive than 2D ones, but 2D video cameras are much easier to use and faster to configure and set-up. On the contrary, using a 3D acquisition system, the detection and tracking of people could be easier and more accurate than using 2D acquisition system.

3.2. Indoor vs outdoor

The environmental conditions are the second aspect to take into account during the design of a new system for pedestrian detection and tracking. In fact, illuminating conditions, as well as the variability of subjects in the scene are critical aspects to be considered. Since outdoor environments generally have tricky boundary conditions which need performing algorithms for their handling, the most of literature on pedestrian detection and tracking systems concerns with applications for outdoor. On the other hand, since the conditions in indoor environments are more controllable, there are applications which show robust and performing human tracking.

Excluding extremely variable indoor crowded places, such as airports, that are considered the same as outdoor, the most of works on pedestrian detection in indoor areas use external markers to detect a moving object inside the scene [21,90,100,104,106–109].

Visual markers are useful easy-to-locate tools and thus allow an easier tracking within a video stream in a controlled environment. Thanks to their non-invasive nature, they could be easily applied on the object to be tracked into a scene. Both the detection and tracking systems are simplified since common RGB cameras could be used to acquire the scene; in particular, high performance pattern matching systems may be used to find the marker inside each frame.

The use of markers in indoor environment may have multiple aims; for example, while in [104], Mehner et al. placed a marker on each person’s head that could be recorded in the scene and have used it to track the different trajectories for subsequent analyses, Naseer et al. describe a system to follow pedestrians using a quadcopter and use markers to support the motion of the quadcopter itself [90]. In particular, the authors set up two cameras,
first one for determining the 3D position of the UAV based on markers placed on the ceiling of a controlled room and second, a depth camera, for detecting a person in the 3D space. The image resulting from the depth camera is then warped, based on the calculated 3D position.

Considering scenarios of human tracking, marker-based systems are highly recommended in applications where the body position needs to be quickly and accurately tracked, while the human skeleton makes unpredictable and complicated motion trajectory. In addition, cluttered scenes, or varied lighting, most likely distract visual attention from the real position of a marker. Given these problems, visual marker-based tracking is preferable.

In these circumstances, simple human tracking is not sufficient; in fact, some applications require to detect and track single body parts, especially in applications within the analysis domain [21,100]. In order to reach high levels of accuracy in human body tracking, it could be necessary to change the marker technology and, consequently, the acquisition system. For example, in [106–109] the authors have used multiple IR cameras with specific visual markers to track human body parts with high accuracy for specific tasks.

From a technical point of view, marker-based tracking systems are easy to implement since, as already stated, markers are employed in controlled environments in terms of lighting and field of view. Scientific community, instead, spent much time to study and implement human detection and tracking system that do not use any kind of marker. In fact, regarding indoor detection, literature reports some recent works describing application were human tracking is performed in indoor environments without any adoption of markers [135,136].

Generally, outdoor pedestrian detection is a more difficult task than indoor one, since the external environment is generally influenced by so many variables and the scenes to be acquired are completely unpredictable. Recent literature contains several works dealing with marker-based human tracking, the most of which make use of drones [51,65,77–79,91,94,96,97,101–103,105,120–123,126,139,140]. In some cases, they do not use external markers to detect and track a human. In [101], the authors have developed a 3D object following system based on visual information acquired from the UAV camera; in particular, the authors recognize a specific object placed into the scene and use this information to control the movement of a drone. In [105], Vasconcelos et al. have used shirts with markers to track a person with a drone; in particular, the authors have developed a “behavioral marker” composed of two different parts: the first one, which is constant, is used for detection and tracking processes; the second part, instead, is variable and is used to adapt the drone behavior to the specific recognized person so that the UAV is able to know which person is targeting and following.

4. Computer vision methods for pedestrian detection

The initial approach for detection and tracking of moving objects into a video flow acquired by a static camera consisted in the Background Subtraction (BS); this technique allows the detection and distinction of moving objects inside a scene using an appropriate background model [48]. Even though algorithms based on BS are quite simple to implement, this approach is not robust to illumination variability, dynamic background, shadows or noise limiting its usage mostly in controlled environments [139–141].

In recent years, a huge number of algorithms have been developed and tested to perform human detection and tracking, but the most of them are based on the following approaches for features extraction and detection:

- **Histograms of oriented gradients [56]:** this method is based on the idea that local object (human or not) appearance and shape can often be characterized considering local intensity gradients or edge directions distribution (Fig. 7). Each video frame is divided into small regions and a local 1-D histogram of gradient directions or edge orientations over the pixels of the block is computed. An improved version of this algorithm, which is able to handle with problems related to illumination or shadowing, is used for the normalization of the histograms considering a group of smaller blocks. In both cases, each generated histogram is considered as image representation and a cascade of classifiers is used to discriminate each sub-region.

- **Haar-like features [50]:** with this approach, the wavelet representation is used to capture the structural similarities between various instances of the class of humans [142]. In particular, 2-dimensional Haar wavelets include basis functions which capture change in intensity along the horizontal, vertical and diagonals (or corners) directions (Fig. 8). As in previous case, each representation is used as input to a classifier. Improved versions of the algorithm are applied to support multi-scale detection.

- **Viola–Jones features [54]:** this approach is an extended version of the rectangle filters presented by Viola and Jones for the static face detection [143,144]; in particular, this approach considers particular filters based on Haar wavelets. In this case, the proposed approach take into account both motion and intensity information even considering sequences of frames.

- **Texture [145]:** features extraction from texture is a quite simple approach and consists in the elaboration of its distribution in the image; in literature several works dealing with textural features extraction could be found [59,143,146]. On the contrary, the classification of pedestrian considering textural features only is a challenging problem due to the high variability of classes to be considered, e.g. pedestrians variations due to clothing and varying lighting conditions. In order to avoid this, textural features are generally used in combination with other kinds of features, such as shape, colour and others.

- **Local Binary Pattern (LBP) [68]:** this technique allows to describe images based on their texture by opportuneely considering the neighborhood of each pixel [147]. LBP approach have become very popular due to its robustness against variations in pose or illumination than other methods. As reported in [68], LBP feature vectors are very often used in combination with HOG features to reach higher performance in pedestrian detection.

In the following paragraph, more details about pedestrian detection will be given related to some innovative works making use of these algorithms or their variations.

4.1. Pedestrian detection and tracking

In this paragraph, the most important works dealing with the task of pedestrian detection and tracking are investigated. In Table 1, the performance of each detector are reported in terms of Log-Average Miss Rate (MR) on the most common benchmark databases, namely Inria [56] and Caltech [25,148], along with details about both the detector and the classifier family used in each work.

In [54], Viola et al. describe a pedestrian detection system that integrates image intensity information with motion details; in particular, the authors combined Haar-like features with motion information that were computed considering two consecutive frames in a video sequence. The authors applied the face detector described in [143] and [144] to the pedestrian detection problem, but their results on the benchmark databases show an high Log-Average Miss Rate. Classification was performed considering a sequence of AdaBoost classifiers.
In [56], Dalal and Triggs studied the question of feature sets for robust visual object recognition, introducing the Histogram of Oriented Gradient (HOG) features. After reviewing existing edge and gradient based descriptors, the authors showed experimentally that grids of HOG descriptors significantly outperform existing feature sets for human detection. A linear SVM was adopted for human detection and classification; Gaussian SVM was explored too, but run-time result does not perform better than using linear SVM.

In [60], the authors addressed the problem of detecting pedestrians in static images introducing a set of features called "Shapelet". These are a combination of low-level features, which consisted primarily in the gradient responses in images, and then in a set of features automatically learned using an AdaBoost classifier. Finally, another AdaBoost classifier was trained to discriminate between pedestrian and non-pedestrian using Shapelet features as input. The reported results show that the developed approach performs better on Caltech database than on Inria one.

In [63], Maji et al. discussed that it is possible to build histogram intersection kernel SVMs (IKSVMs) with a logarithmic run time complexity considering the number of support vectors as opposed to linear used as standard approach. The authors introduced a variant of HOG features based on a multi-level version of HOG descriptors. They showed that by pre-computing auxiliary tables, it was possible to design an approximate classifier with constant

---

**Table 1**

<table>
<thead>
<tr>
<th>Detector</th>
<th>Training set</th>
<th>Classifier</th>
<th>Test set</th>
<th>Log-Average Miss Rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informed Haar [93]</td>
<td>Caltech</td>
<td>AdaBoost</td>
<td>Caltech</td>
<td>34.60</td>
</tr>
<tr>
<td>Informed Haar [93]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>14.43</td>
</tr>
<tr>
<td>VJ [54]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Caltech</td>
<td>94.73</td>
</tr>
<tr>
<td>VJ [54]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>72.48</td>
</tr>
<tr>
<td>HOG [56]</td>
<td>Inria</td>
<td>linear SVM</td>
<td>Caltech</td>
<td>68.46</td>
</tr>
<tr>
<td>HOG [56]</td>
<td>Inria</td>
<td>linear SVM</td>
<td>Inria</td>
<td>45.98</td>
</tr>
<tr>
<td>Shapelet [60]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Caltech</td>
<td>91.37</td>
</tr>
<tr>
<td>Shapelet [60]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>81.70</td>
</tr>
<tr>
<td>MultiFtr + CSS [71]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Caltech</td>
<td>60.89</td>
</tr>
<tr>
<td>MultiFtr + CSS [71]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>24.74</td>
</tr>
<tr>
<td>MultiFtr + Motion [71]</td>
<td>TUD-Motion</td>
<td>linear SVM</td>
<td>Caltech</td>
<td>50.88</td>
</tr>
<tr>
<td>HikSvm [63]</td>
<td>Inria</td>
<td>HIK SVM</td>
<td>Caltech</td>
<td>73.39</td>
</tr>
<tr>
<td>HikSvm [63]</td>
<td>Inria</td>
<td>HIK SVM</td>
<td>Inria</td>
<td>42.80</td>
</tr>
<tr>
<td>HogLhp [68]</td>
<td>Inria</td>
<td>linear SVM</td>
<td>Caltech</td>
<td>67.77</td>
</tr>
<tr>
<td>HogLhp [68]</td>
<td>Inria</td>
<td>linear SVM</td>
<td>Inria</td>
<td>39.10</td>
</tr>
<tr>
<td>LatSvm-V1 [64]</td>
<td>Pascal</td>
<td>latent SVM</td>
<td>Caltech</td>
<td>79.78</td>
</tr>
<tr>
<td>LatSvm-V1 [64]</td>
<td>Pascal</td>
<td>latent SVM</td>
<td>Inria</td>
<td>43.83</td>
</tr>
<tr>
<td>LatSvm-V2 [72]</td>
<td>Inria</td>
<td>latent SVM</td>
<td>Caltech</td>
<td>63.26</td>
</tr>
<tr>
<td>LatSvm-V2 [72]</td>
<td>Inria</td>
<td>latent SVM</td>
<td>Inria</td>
<td>19.96</td>
</tr>
<tr>
<td>ChnFtrs [67]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Caltech</td>
<td>56.34</td>
</tr>
<tr>
<td>ChnFtrs [67]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>22.18</td>
</tr>
<tr>
<td>FeatSynth [75]</td>
<td>Inria</td>
<td>linear SVM</td>
<td>Caltech</td>
<td>60.16</td>
</tr>
<tr>
<td>FeatSynth [75]</td>
<td>Inria</td>
<td>linear SVM</td>
<td>Inria</td>
<td>30.88</td>
</tr>
<tr>
<td>MultiRec [73]</td>
<td>Caltech</td>
<td>latent SVM</td>
<td>Caltech</td>
<td>48.45</td>
</tr>
<tr>
<td>CrossTalk [80]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Caltech</td>
<td>53.88</td>
</tr>
<tr>
<td>CrossTalk [80]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>18.98</td>
</tr>
<tr>
<td>VeryFast [81]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>15.96</td>
</tr>
<tr>
<td>SketchTokens [84]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>13.32</td>
</tr>
<tr>
<td>Roeri [85]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Caltech</td>
<td>48.35</td>
</tr>
<tr>
<td>Roeri [85]</td>
<td>Inria</td>
<td>AdaBoost</td>
<td>Inria</td>
<td>13.53</td>
</tr>
<tr>
<td>AFS-Geo [86]</td>
<td>Inria</td>
<td>linear SVM</td>
<td>Caltech</td>
<td>66.76</td>
</tr>
<tr>
<td>DBN-Isol [82]</td>
<td>Inria</td>
<td>DeepNet</td>
<td>Caltech</td>
<td>53.14</td>
</tr>
<tr>
<td>DBN-Mix [88]</td>
<td>Inria</td>
<td>DeepNet</td>
<td>Caltech</td>
<td>48.22</td>
</tr>
<tr>
<td>ACF+SDT [85]</td>
<td>Caltech</td>
<td>AdaBoost</td>
<td>Caltech</td>
<td>37.34</td>
</tr>
</tbody>
</table>
runtime and space requirements, independent of the number of support vectors, with negligible loss in classification accuracy on various tasks.

In [64], Felzenszwalb et al. described an approach based on part model for object detection. The authors evaluated HOG features at different levels of resolution leading to "HOG features pyramid", thus allowing the detection of parts that could be moved respect to the detection window. The authors combined a margin-sensitive approach for data mining hard negative examples with a formalism called latent SVM which leads to a non-convex training problem. However, a latent SVM is semi-convex and the training problem becomes convex, once latent information was specified for the positive examples. In [72], Felzenszwalb et al. reduce the dimensionality of the dataset used in [64] through the PCA algorithm. An improved version of the multi-scale detection, together with PCA for dimensionality reduction, led to an improvement of the performance on both the benchmark test sets Inria and Caltech.

In [68], Wang et al. proposed a novel human detection approach capable of handling partial occlusion. In details, a new feature set was introduced considering HOG and LBP features. In order to handle partial occlusions, two detectors were combined: the first is performed globally on the image, while the second (part detector) is executed in ambiguous areas to refine the detection. For each ambiguous scanning window, an occlusion likelihood map was constructed by using the response of each block of the HOG feature to the global detector. The occlusion likelihood map was then segmented by Meanshift approach [149]. The segmented portion of the window with a majority of negative response is inferred as an occluded region. Thanks to this this approach, based on the augmented HOG-LBP feature and the global part occlusion handling method, they achieved very high levels of detection rates considering linear SVM classifiers.

In [67], Dollar et al. studied the performance of "integral channel features" for image classification task, focusing in particular on pedestrian detection. In details, multiple representation of the same input image could be computed applying linear and non-linear transformations. Considering the features extracted from each representation, such as local sums, histograms, and Haar features and their different generalizations, the integral image is then computed [150] and used as input in the classification step. Performance was tested considering three different classifiers: AdaBoost, RealBoost and LogitBoost. In [80], Dollar et al. also investigated the correlations between detector responses at nearby location and scales in an application where cascades help to make sliding windows object detection fast, nevertheless, computational demands remain prohibitive. In particular, the authors selected a restricted subset of features from the group reported in [67], focusing their work on a low-level optimization, leading to an improvement at both compile and run time as well.

In [71], Walk et al. showed that motion features derived from optical flow, if implemented correctly, yield substantial improvements on image sequences, even in presence of low-quality video sequences. The authors introduce a novel feature which called "CSS" based on the self-similarity of low-level features capturing pairwise statistics of specially localized colour distributions. Subsequently, the authors firstly evaluate performance of classification coupling HOG features with CSS; then, to the previous group of features, motion information was added and then computed as a variant of Histogram Of Flows (HOF) algorithm proposed by Dalal et al. [151]. The latter approach consistently improved the detection performance both for static images and video sequences, across the two different datasets. In combination with HOG, these two features outperform the state-of-the-art by up to 20%. In [71], Linear SVM was used to classify and evaluate the performance; then, a variant of AdaBoost algorithm (MLPBoost) was tested also on Caltech dataset.

In [75], Bar-Hillel et al. introduced a new approach for learning part based object detection through feature synthesis in pedestrian detection task. The authors considered different families of features, e.g. HOG, LocalMax or Sift, and for each iteration of their algorithm, a subset of features was used in the generation process, by using pruning strategy as well. In details, the described method consists of an iterative process of feature generation and pruning, in which basic part-based features are developed into a feature hierarchy using operators for part localization, part refining and part combination. Then, feature pruning was performed by using a new features selection algorithm for linear SVM, namely Predictive Feature Selection (PFS), based on weight prediction.

In [73], Park et al. described a multi-resolution model that acts as a deformable part-based model when scoring large instances and a rigid template with scoring small instances. Substantially, the authors demonstrated the necessity to extract features and classify at multi-resolution stages to avoid miss-detection. As in [64], latent SVM were used in the classification step, and the authors demonstrated impressive results on the Caltech Pedestrian benchmark.

In [81], Benenson et al. presented a new pedestrian detector that efficiently handles different scales avoiding the resize of input images; by transferring computation from test time to training time, detection speed was optimized and improved. When processing monocular images, the system provides high quality detections at 50 fps. The authors also proposed a new method for exploiting geometric context extracted from stereo images with high fps by using a CPU+GPU machine.

In [84], Lim et al. proposed a novel approach to both learning and detecting local contour-based representations for mid-level features. The features, called sketch tokens, are learned using supervised mid-level information in the form of hand drawn contours in images. Combining sketch tokens features with the integral image approach proposed by Dollar et al. [67,80] the authors reported a slight improvement in the pedestrian detection approach, with a classification achieved by Random Forest Algorithm.

In [85], Benenson et al. revisited the core assumptions of HOG+SVM algorithm and showed that, by properly designing the feature pooling, feature selection, pre-processing, and training methods, it is possible to reach high performance in pedestrian detection. The authors described an approach based on the multi-scale model generation introduced in [81] but, in contrast with the algorithm proposed by Dalal et al. [56], the windows considered for features extraction selected during learning, were composed by irregular patterns.

In [86], Levi et al. presented a new part-based object detection algorithm with hundreds of parts performing real-time detection based on the approach proposed in [75] by Bar et al. However, due to their high computational demands part-based methods are
limited to several parts only and are too slow for practical real-time implementation. The authors proposed the Accelerated Feature Synthesis (AFS) algorithm and, in order to reduce the number of locations searched for each part, introduced an algorithm for approximate nearest neighbor (KDFerns), to compare each image location to only a subset of the model parts. Candidate part locations for a specific part are then further reduced by using spatial inhibition, and using an object-level " coarse-to-fine" strategy. Linear SVM was used in the classification step.

In [89], Park et al. introduced a combined approach for motion information extraction from video sequences. Prior to features extraction, the authors performed a weak motion stabilization by considering both camera and object motion, and at the same time preserved non-rigid motion that provided useful information for the recognition task. The authors also described a combined approach that used coarse-scale flow and fine-scale temporal difference features and used AdaBoost for classification.

In [93], Zhang et al. proposed a pedestrian detection algorithm introducing several efficient features based on Haar wavelets, called "Compact Features". In the reported work, the authors assume that pedestrians, in the most of cases, show a recurrent behavior, or rather the first visible part of each pedestrian is the upper-right (head and right part of the shoulder). Following this approach, the authors employed a statistical model of the up-right human body where the head, the upper body, and the lower body are treated as separated parts and, in this way, partial occlusions were handled allowing to reach high performances with an occlusion higher than 35%. The classifier used in [93] was AdaBoost.

More recently, Cao et al. proposed a pedestrian detection algorithm considering a set of features appearance constancy and shape symmetry, called NNFE, constituted by both Non-Neighboring (NNF) and Neighboring Features (NF) [152]. The proposed approach have been tested on Caltech dataset reporting good performances compared with state-of-the-art methods.

4.1.1. 2D vs 3D

Features extraction consists in different methods that transform one or more input images into a reduced representation that could be used as input to classifiers. Thanks to different strategies, it is also possible to reduce the dimensionality of these patterns allowing a faster and more accurate classification [153–158]. For the aim of pedestrian detection, the extraction of features is a fundamental task and it is independent from the acquisition technology. In such a kind of applications, pedestrian detection is performed considering RGB images in both 2D and 3D applications. The additional information coming from the third dimension, regardless of stereo-vision or depth cameras, is especially used for tracking pedestrians, allowing to keep track of their position in a 3D space; this kind of approach is used in applications of PPSs where the mutual positions of pedestrians and the moving object (i.e. an autonomous car) are of fundamental importance to correctly control the object.

5. Machine learning techniques for pedestrian detection

Data mining techniques, including machine learning, have been used to learn hidden information in data in order to train automatic systems for decision making processes in several domains [159,160].

Since an acquired scene may contain several kinds of objects candidate for tracking, image processing techniques often fail to filter out background and/or objects of other classes; thus, machine learning methods may help in discriminating pedestrian from other classes of objects in the scene. According to the workflows introduced in Section 1, both traditional approaches and deep learning strategies are used for classification.

Regarding traditional approaches applied to pedestrian detection, the dataset created from the features extracted after the processing of input images influence the design of the classification strategy [161]. In particular, from the input dataset point of view, several algorithms for dataset processing, such as normalization or dimensionality reduction, have been developed, and are applied to improve the classification performance [162–164]. Even from the classifiers point of view, there are several classification algorithms used to perform pedestrian detection, the most of which consist in supervised approach, such as Support Vector Machine (SVM), Artificial Neural Network (ANN), or Boosting algorithms.

Regarding Deep Learning strategies, instead, the design of deep classifiers following the workflow reported in Fig. 3, such as Convolutional Neural Networks, the main task to address is the design of the network topology. In fact, the number of hidden layers strongly influences the network performance in terms of both classification accuracy and execution time. Although a universal strategy to design a good classifier does not exists, the tradeoff between classification performance and training time of the classifier should lead the design of the topology. Specifically, a number of layers too low reduce the training time but the model could too simple for the classification task; on the contrary, a number of layers too high could lead to the classifier overfitting on training data reducing the classifier performance on new data.

In the following sections, the traditional approaches of machine learning will be discussed, analysing the most common architectures employed for pedestrian classification. Furthermore, Deep Learning strategies will be introduced and discussed focusing on deep structures applied on the pedestrian detection task.

5.1. The traditional approaches

As could be seen in Table 1, which reports the performance of the algorithms discussed in Section 4.1, almost all the considered works for pedestrian detection and tracking use simple classifiers, such as Support Vectors Machines (SVMs) or boost families.

Since the pedestrian detection and tracking tasks have an high computational cost, especially in real time applications, very often in literature are presented classifier models with low complexity. Linear SVM and weak decision trees with low depths boosted to speed up the learning phase are the most used, since they can lead to lighter decision processes making the image processing part the most important in the decisional process.

Even if the SVM’s design, in terms of complexity, is an automatic procedure for selecting Support Vectors [165], we present the Artificial Neural Networks (ANNs) performances on the mentioned benchmark databases [166–171], as the case of Zhao et al. that developed a stereo-system for pedestrian classification [172].

In [145], Gravila and Munder developed a system, called PROTECTOR, constituted by several processing modules, one of which consists in a neural model that classify pedestrian based on textural features extracted from each video frame.

In fact, thanks to suitable optimization strategies [173–176] it is possible to find the optimal topology for an ANN to classify two or more classes in the best way and, by using a multi-objective algorithm, the topology could be optimized, thus allowing a faster classification in various research topic [177–185].

5.2. The deep learning approaches

Recent researches in Artificial Intelligence (AI) led to the spread of modern techniques of machine learning based in deep structures, as reported in novel and innovative works, [110,111,114,117,118,186–188].

Deep Learning strategies have been used for automatic object detection and images segmentation and classification applications.
The most diffused DL architectures are Convolutional Neural Networks, which are able to classify images into several categories, automatically learning features through convolutional layers that combine multiple non-linear processes.

Since the training of a CNN is very time and computation resource consuming, two different approaches have been found in literature for CNNs: (i) Transfer Learning, that allow to “re-train” a pre-trained model on different categories (e.g. use AlexNet to discriminate among different kind of tumours); (ii) Feature Extractors, as CNNs are constituted by several convolutional layers which create different layer of features representations, it is possible to catch each layer output and use it as input to simpler classifiers, such as SVM or ANN.

Based on Convolutional Neural Networks, these approaches have the ability to learn effective hierarchical feature representations that characterize the typical variations observed in visual data, including images and video, which make them very well suited for the most of visual classification tasks.

For pedestrian detection, Szarvas et al. used CNN to classify pedestrian in images [189]. The authors compared their approach to classical SVM approach with Haar features obtaining higher levels of accuracy. Then, the CNN was used as features extractor and the computed descriptors were used as input to a Gaussian-SVM classifier and the reported results were increased respect to the CNN approach for classification.

Automatic features extraction in also used in the work by Zhang et al., where the authors used faster r-cnn for pedestrian detection [190]. The developed system was composed of two cascaded subsystems: the first was deputy to detect candidate regions in the image that could contain a pedestrian; the second sub-system, instead, was a Boosted Forest classifier for the pedestrian classification [191,192].

Recently, Li et al. have used neural features by applying fully convolutional neural networks as features extractors [186]. In details, the authors have tested and compared the performance of AdaBoost classifiers by using input extracted at different levels from the network. The reported results on benchmark datasets are very promising (Log-Average Miss Rate about 20%) if compared with those reported in Table 1.

Since occlusions are one of the most discussed problems in literature [25,26,68,82,93,149,193], deep learning allowed to strengthen the detection of single parts in order to find and correctly classify occluded pedestrian [82,88,115,116].

Ouyang and Wang presented a probabilistic pedestrian detection framework to solve the issue related on the inaccurate scores of part detectors when there are occlusions or large deformations [82]. In this framework, a deformable part-based model was used to obtain the scores of part detectors and the visibilities of parts were modeled as hidden variables. In the proposed work, a discriminative deep model based on Restricted Boltzmann Machine (RBM) building blocks was used for learning the visibility relationship among overlapping parts at multiple layers. Experimental results on benchmark datasets showed the effectiveness of the proposed approach. An improved version of the proposed algorithm is reported in [88] where Ouyang et al. proposed a mutual visibility deep model that jointly estimates the visibility statuses of overlapping pedestrians using Gaussian Mixture Model (GMM). The visibility relationship among pedestrians was learned from the deep model for recognizing co-existing pedestrians. Experimental results showed that the mutual visibility deep model effectively improved the pedestrian detection results. In [116], the main idea is to construct multi-parts detectors that covers several scales of different body parts and automatically choose important parts for occlusion handling. At the training stage, each part detector is learned by CNN fine-tuning approach, using a CNN pre-trained on ImageNet Database [27]. At the testing stage, a shifting handling method within a CNN is designed. This method handles the problem that positive proposal windows usually shift away from their corresponding ground truth bounding boxes. Moreover, the part selection is determined by data and the effectiveness of the part pool can be fully explored.

Human body pose recognition is also a well-suited task for DL approaches [112,113,115,119,194,195]. Human body pose recognition in video is a long-standing problem in computer vision with a wide range of applications. However, body pose recognition remains a challenging problem due to the high dimensionality of the input data and the high variability of possible body poses. As reported in the previous section, traditional computer vision-based approaches are mostly based on appearance cues such as textures, edges, colour histograms, foreground silhouettes or hand-crafted local features (such as histogram of gradients (HOG) [56]) rather than motion-based features. Alternatively, psychophysical experiments have shown that motion is a powerful visual cue capable to extract high-level information, including articulated pose [196]. In particular, a combination of hand-crafted features and DL classifier may be a good approach to estimate human pose. For example, in [112], it is shown that deep learning is able to successfully incorporate both RGB and motion features for the task of human body pose detection in video.

However, to estimate human body pose, deep learning approach to predict a single class label per image has to be supported by a high resolution semantic segmentation output. To reach this result, Oliveira et al. [119] used the so called “up-convolutional networks” [115,194]; in contrast to usual classification, which contracts the high-resolution input to a low-resolution output, this kind of networks can take an abstract, low-resolution input and predicts a high-resolution output, such as a full-size image. To reach this goal, it is possible to refine the architecture of Long et al. [115] and apply it to human body part segmentation to use it different contexts, such as robotics.

Regarding robotics, human body parts segmentation can be a very valuable tool, especially when it can be applied both indoor and outdoor. For persons who cannot move their upper body, some of the most basic actions, such as drinking water, is rendered impossible without assistance. Robots could identify human body parts, such as hands or harms, and interact with them to perform some of these tasks. Other applications, such as learning from demonstration and human robot handover can also benefit from accurate human part segmentation. For a learning-from-demonstration task, one could take advantage of the high level description of human parts, considering each of them as an explicit mapping between the human and joints of the robot for learning control actions. A robot that needs to hand a tool to its human counterpart must be able to detect where the hands are to perform the task. Human body part segmentation has been considered a very challenging task in computer vision due to the wide variability of the body parts’ appearance, pose and viewpoint; self-occlusion and clothing, also, represents very difficult problems to handle.

In [113] the pose estimation is formulated as a Deep Neural Network (DNN)-based regression problem towards body joints. A cascade of such DNN regressors which results in high precision pose estimates is presented. The considered approach has the advantage of reasoning about pose in a holistic fashion and has a simple but yet powerful formulation which capitalizes on recent advances in Deep Learning. DNNs have shown outstanding performance on visual classification tasks [27] and more recently on object localization [197,198].
6. Discussion and future trends

In the last decades, pedestrian detection and tracking systems gained a considerable importance thanks to their versatility use. The study and development of systems able to automatically interact with moving humans have introduced the need to increase the performance of human detection and, at the same time, improve the run-time performance.

A deep analysis of the results reported in Table 1 is necessary. Table 2 shows different metrics related to the performances of pedestrian detection systems on Inria and Caltech datasets (mean, standard deviation, min and max values). As could be seen, the mean value of Log-Average Miss Rate is significantly higher for detectors on Caltech dataset than on Inria one (lower is better) as shown in Fig. 9 \((p \leq 0.005)\). Conversely, the classification performances are not specifically related to the considered classifier family (Fig. 10), even though AdaBoost performs better than SVM on average, as reported in Table 3.

This important result highlights that the implemented classifiers perform better on static images classification than videos containing more noisy classes that could disturb the pedestrian detection. As demonstrated by the works that have shown the best performances on Caltech dataset [89,93], the detection and classification of pedestrian in videos has to be supported by sets of features that take into account motion information too.

A further analysis have been conducted by analysing the performance by grouping classifiers and test sets; in detail, four groups have been created which were: G1 - AdaBoost classifier on Inria; G2 - AdaBoost classifier on Caltech; G3 - SVM classifier on Inria; G4 - SVM classifier on Caltech. The results reported in Table 4 confirm the higher capabilities of classifiers to discriminate pedestrians on the static images from Inria dataset, regardless of the considered classifier families (Fig. 11).

The introduction of Deep Learning architectures, as well as the accessibility of cheaper but more powerful computers, led the scientific community to study more performing systems for two main reasons: (i) DL architectures may help to design more informative sets of features; (ii) DL architectures performance at execution time are faster than traditional models of machine learning.

The computer vision systems adopted to perform pedestrian detection differ based on the acquisition sensor; in particular, 2D sensors limit the task of pedestrian detection to a bi-dimensional space. On the other side, stereo-cameras and depth sensors are able to track pedestrians in the 3D space.

Moreover, some of the works reported in this survey make use of markers; the kind of marker, as well as the aim of each work and the desired level of accuracy to be reach, strongly influence the computer-vision system for images acquisition.

Since human tracking is applied in multiple scenarios, the literature reports a very large variety of configurations for the vision system but, at the same time, the classifiers used to discriminate humans, or pedestrians, among the multitude of objects in the scenes, are limited to the simpler classifiers, such as SVM or decision trees. Artificial Neural Networks are quite used, but very limi-
Fig. 11. Box plot of the Log-Average considering four couples of Classifier and Test Set. G1 - AdaBoost classifier on Inria; G2 - AdaBoost classifier on Caltech; G3 - SVM classifier on Inria; G4 - SVM classifier on Caltech (*p ≤ 0.05  **p ≤ 0.01  ***p ≤ 0.001).

ated respect to the previous models, besides recent works demonstrated their versatility in different domains [199–202].

The most difficult step in the design of pedestrian detection system concerns with the features extraction, as it is necessary to extract powerful descriptor that have to help to discriminate pedestrian. Thanks to the introduction on Deep Learning structures, the previous step could be bypassed since deep architecture, such as CNNs, could automatically create their own representation of features.

Following the previous conclusions, in order to design novel applications for pedestrian detection, several aspects had to be considered. First, it is necessary to design the desired degree of accuracy to be reached; this influences the technology of the acquisition system: an RGB camera could be sufficient to detect pedestrian in 2D space, using background subtraction if it is possible create a simple model of the background; or feature-based approaches to classify pedestrians are to be considered: linear and non-linear classifiers, such as SVM or ANNs, may be considered in cascade to the previous step, or a Deep Learning strategy may be implemented using, for example, Convolutional Neural Networks. In this latter case, there is no need to extract features, but an efficient strategy for objects detection in the image has to be implemented. In any case, a powerful approach may consist in the combination of the two proposed strategies; in details, deep architectures may be used to extract features (even at different levels of abstraction) to be used as input to simple learner for pedestrian classification.

The necessity to track pedestrian in the 3D space imposes the use of depth cameras, or stereo-cameras. This combination is necessary when both the presence and the movements of pedestrian control one or more automatic machine in the real world, such as drones or autonomous vehicles. In some cases, for example in controlled indoor environment, it is necessary “to help” the tracking system with markers. These scenarios are the most common approach in applications that involve drones (even if scientific community is taking the lead of marker-less strategies) or research purposes, such as the study of crowded places tracking and analysing the pedestrian trajectories.

For clinical purposes, instead, the use of multiple markers placed on human body is strongly recommended to accurately track human body parts. In fact, there are only few works that track people without any kind of support for clinical purposes.

6.1. Future works

The future of pedestrian detection concerns with the improvement of performance of both detectors and classifiers. In fact, improving the speed of pedestrian detection has been an active area in recent years. For example, in [81], Benenson et al. proposed a method reaching speeds of 100–135 FPS for detection in a 480 × 640 image, although the levels of accuracy are still low. Other researchers have focused specifically on speeding up Deep Neural Networks [198,203,204], but with no real-time solutions. In [205], Angelova et al. presented a new real-time approach to object detection that exploits the efficiency of cascade classifiers with the accuracy of deep neural networks.

Excellent performance of Deep Networks in classification tasks are found in literature, and their ability to operate on raw pixel input without the need to design special features is very appealing. However, deep nets are notoriously slow at inference time. In this work, the authors proposed an approach that cascades deep nets and fast features, that is both very fast and very accurate. They applied it to the challenging task of pedestrian detection. Their algorithm runs in real-time at 15 fps. The resulting approach achieves a 26.2% average miss rate on the Caltech Pedestrian detection benchmark, which is competitive with the very best reported results. The importance of pedestrian real-time detection is particularly relevant in advanced driver assistance systems (ADASs), and pedestrian protection systems (PPSs) [76]. As a future work, it could be interesting to find the best trade-off between accuracy and frames per second in different environmental conditions and contexts (e.g. Human-Aware Navigation to detect falls [206]).

7. Conclusion

In this work, a survey on pedestrian detection and tracking system have been presented. Recent adoption of Deep Learning methodologies and in particular of Convolutional Neural Networks for pedestrian detection and tracking deserved a dedicated state-of-the-art survey. The analysed works highlight the need to investigate how modern approaches to pedestrian detection work and a comparison with the features-based approaches on benchmark datasets has to be done.

However, the reported works show encouraging results in automatic pedestrian detection, but further architectures need to be implemented and tested. In particular, for pedestrian detection, the most successful way seems to consist in the combination of Deep Learning with classical Machine Learning models because this seems to imply high levels of accuracy and less computation respect to hand-designed features and classification. Moreover, it will be interesting to compare the performance to this task of optimal ANNs topologies with SVM.

References


Vitoantonio Bevilacqua obtained the Laurea Degree in Electronic Engineering and the Ph.D. in Electrical Engineering from Polytechnic University of Bari, was qualified as Associate Professor of Information Processing Systems and in March 2017 also as Associate Professor of Bioengineering. Currently he is Professor of Human Computer Interaction at the Department of Electrical and Information Engineering of Polytechnic University of Bari where previously also taught Expert Systems, Medical Informatics and Image Processing, and where is the Head of Industrial Informatics Lab. Since 1996 he has been working and investigating in the field of computer vision and image processing, human-machine interaction, bioengineering, machine learning and soft computing (neural networks, evolutionary algorithms, hybrid expert systems, deep learning). The main applications of his research are in medicine, in biometry, in bioinformatics in ambient assisted living and industry. In 2000 he was involved as Visiting Researcher in an EC funded TMR (Trans-Mobility of Researchers) network (ERB FMRX-CT97-0127) called CAMERA (CAD Modeling Environment from Range Images) and worked in Manchester (UK) at UK Robotics Ltd, in the field of geometric feature extraction and 3D objects reconstruction. He has published more than 145 papers in refereed journals, books, international conferences proceedings and chaired several sessions such as Speech Recognition, Biomedical Informatics, Intelligent Image Processing and Bioinformatics in international conferences. On July 2011, he was invited as lecturer at International School on Medical Imaging using Bioinspired and Soft Computing-Miere (Spain) MIBISOC FP7-PEOPLE-ITN-2008. GA N. 238819 where presented his research on Intelligent Tumors Computer Aided Early Diagnosis and Therapy: Neural Network and Genetic Algorithms frameworks. He served as 2009 and 2016 International Conference on Intelligent Computing Program Chair, as 2015 and 2017 Human Machine Interaction Summer School General Chair and as ISMISS 2016 IEEE Italy Section Medical Informatics Summer School Co-Chair.