Pedestrian Attributes and Activity Recognition Using Deep Learning: A Comprehensive Survey

Mohammed Fadhil Asghar [1], Mudhafar Hussein Ali [2], Jumana Waleed [3]

[1,2] Department of Computer Engineering, College of Engineering, Al-Iraqia University, Iraq
[3] Department of Computer Science ,College of Science, University of Diyala, Iraq

Abstract

In recent years, pedestrian attributes and activity recognition has attracted expanding research emphasis due to their considerable research importance and value of application in the intelligent civil and military domains. Owing to the inadequate image or frame quality of inexpensive cameras and the absence of obvious and stable feature information, direction of pedestrian movement, and so on, the complication of pedestrian attributes and activity recognition is expanded. However, with the comprehensive implementation of deep learning techniques, pedestrian attributes and activity recognition has made a substantial advance. This paper is the first of its kind because it combines the recognition of pedestrian attributes and activities separately, and reviews the works on pedestrian attributes and activities using deep learning in relation to datasets. The fundamental concepts, corresponding challenges, and popular solutions are also explained. Furthermore, in this community, metrics of evaluation and concise performance comparisons are given. In the end, the hotspots of the present research and the directions of the future research are summarized.

Keywords:- Pedestrian Attribute (PA) recognition, Pedestrian Activity recognition, Deep learning.

I. INTRODUCTION

In the computer vision area, pedestrian attributes and activity recognition is a successful study topics. Pedestrian attributes (PA) recognition is all about extracting visual qualities from a set of images of people. The detected attributes could be to several classifications, including dress style, footwear, gender, age group, etc. [1]. During pedestrian activity recognition, the mentioned activities may be classified into several categories, such as walking, running, jumping, etc.

Deep learning is a machine learning subfield that focuses on a novel approach to learning representations from data by placing a concentration on learning successively layers of growing meaningful representations. The term "deep learning" doesn't allude to any form of deeper knowledge that may be attained by the method; rather, the "deep" in "deep learning" refers to the notion of successive layers of representations. Deep learning, as it is used currently, often contains tens or even hundreds of consecutive layers of representations, and they're all learned automatically from exposure to training data. Other methods of machine learning, on the other side, tend to concentrate on learning just one or two layers of representations of the data; this kind of learning is frequently referred to as shallow learning. In deep learning, these layered representations are learned using models known as neural networks, which are structured in layers piled on top of one another. [2]

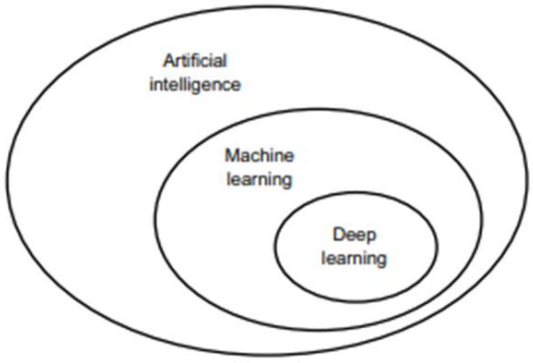


Figure 1: Artificial intelligence, machine learning, and deep learning. [2]

Deep learning is the paradigm that has drastically altered the landscape of artificial intelligence in just a few years. The topic of deep learning is artificial neural networks that contain a lot of layers. The majority of effective applications could be found in visual picture comprehension, as well as audio and text modeling [3]. In PA recognition, there are two major challenges with attribute inference at a great distance: Firstly, Appearance diversity, Due to the varying looks of pedestrian apparel and uncontrolled label multi-factor variations like illumination and the angle of view of the camera, In regard to the same attribute, there exist big intra-class variations between diverse photos. Learning to reveal such attributes needs a considerable number of training examples. Because of inherent data bias, using a single source and limited training data may result in a model that is unrealistic and generalizes poorly to unknown domains. Secondly, Appearance ambiguity and far-view attribute detection are incredibly difficult problems due to inherent visual ambiguity and the lower quality of visual data extracted from the field of a distant view. A single Image May Only include a little Tens of imaging Pixels, of which only a small fraction are recognizable for attribute categorization. Frequently, Obstacles or other pedestrians have obscured parts of the body, which makes it more difficult to extract pertinent features for inference [4].

In addition to the aforementioned, as represented in [5], the following is a list of difficult factors that may clearly affect the final recognition performance: Firstly, Multi-views, or images captured from many camera perspectives, provide viewpoint issues for a variety of computer vision applications. Since the human body is not a solid object, it is more difficult to recognize a person's attributes. When the body is partially concealed by another person or item. Secondly, the Occlusion and Person qualities are more difficult to distinguish. Due to the fact that the Pixel values supplied by obstructed sections may confuse the model and lead to imprecise predictions. Thirdly, Imbalanced attribute distribution, since each human has various attributes, unbalanced data distribution is the outcome of a variable number of different attributes. Fourthly, Low resolution, since high-quality cameras are pricey rather so the image resolution is low. Fifthly, Illumination, Images may be captured at any moment during twenty-four hours. Consequently, at various times, the lighting conditions vary. The shadow can also be captured in the person's photos, and nighttime photos may be completely ineffectual. And finally, Blur, as someone is moving, Pictures captured by a camera could be blurry. Consequently, recognizing attributes under these scenarios is as well a very severe task. Figure 2 illustrates the challenges of recognizing PA.

Figure 2: The challenges of pedestrian attribute (PA) recognition

Regarding survey or review papers, we observed that there are no survey or review papers in the area of pedestrian activity recognition and just a few survey papers in the area of PA recognition. Wang et al. [5] presented a survey paper in which the existing works of PA recognition using deep learning or traditional methods networks were explained. This survey paper an overview of PA recognition, including the fundamental concepts underlying the development of PA as well as the obstacles that come along with its implementation. And studied significant answers to this challenge from a total of eight different points of view. Then, compare deep learning and conventional model-based PA recognition techniques. Then, it introduced the benchmark dataset/(s), and evaluation metrics, and offered a terse comparison of performance. X. Chen et al. [6] also presented a survey paper that presented the fundamental concepts of PA recognition, and provide the current work of PA recognition utilizing deep learning, moreover, presented the commonly utilized datasets and evaluation criteria. A comparison between the presented survey and the other survey papers in the field of pedestrian attributes and activity recognition is demonstrated in Table 1.

Table 1: Comparison with other survey papers.

Ref.	Year	Study type	Machine Learning	Deep Learning	Field
[5]	2022	Survey	Yes	Yes	PA Recognition
[6]	2022	Survey	No	Yes	PA Recognition
Our Survey	2022	Survey	No	Yes	Pedestrian Attributes and Activity Recognition

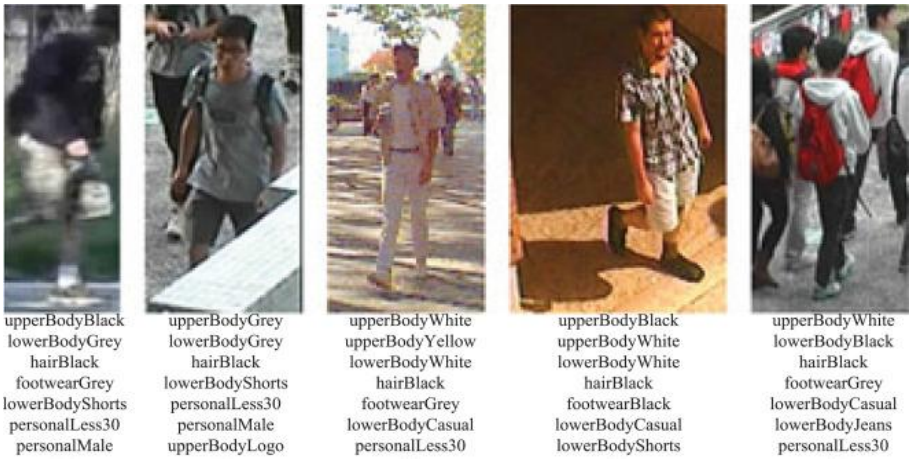
II. PEDESTRIAN ATTRIBUTES RECOGNITION

Attribute recognition for pedestrians seeks to predict, from a list of predefined ones, an attributes category that describes the individual's attributes [7]. In contrast to other computer vision challenges, the dataset annotation for PA recognition contains multiple designations at various levels, for instance, hat, glass, and hair color & style are all low-level characteristics that related to various parts of the images. But some attributes, like gender, sexual orientation, and age, are abstract ideas that don't correspond to specific places. We call these high-level attributes. Also, the factors of context and environment like the parts of the body, viewpoint, occlusion, and can have a big effect on recognizing human attributes [5]. Some datasets include notes about perspective, components' bounding boxes, and occlusion, which makes this research easier.

1) PEdes Trian Attribute (PETA) dataset

PEdes Trian Attribute (PETA) dataset [4] is attribute dataset for pedestrians including more than 60 characteristics on 19000 photos. Robust attribute detectors with excellent generalization performance may be learned more easily with a big dataset. PETA dataset was produced by selecting and organizing 10 publically accessible small-scale datasets. The resolution varies from 17×39 to 169×365 pixels. The organization of these datasets is not easy, erroneous photos or duplicate copies were meticulously eliminated from each dataset. In addition, 61 binary and 4 multi-class attributes have been labeled on each picture. Binary attributes include a group of important characteristics, inclusive of the person's physical characteristics (such as gender and age), look (such as hairstyle), and dress choices (such as casual or formal) on the upper and lower bodies, and accessories. The four things that apply to more than one class include eleven basic color naming for hair, upper body, lower-body, and foot wear. If the ratio of the biggest to the smallest category is no more than 20:1 the binary attribute distribution is deemed to be balanced. Intrinsically, from sixty-one binary attributes, thirty-one consider balanced. The distribution of some features with sample images is shown in Figure 3. And some images with their corresponding attributes.

(a)



(b)

Figure 3: The PETA dataset; (a) Examples and distributions of certain attributes (blue: positive, orange: negative), (b) Some image examples and corresponding attributes [4].

Lately, deeply learning had led into tremendous advances in the automated extraction of features from multi-layer nonlinear-transformations. In this configuration, the PETA dataset is frequently used via recently existing works.

In 2018, Wenhua Fang et al. [7] proposed a combined Using convolutional neural networks and a hierarchical multi-task learning model, it is possible to learn how different attributes are related to each other so that best recognize pedestrians in still images using CNN. The fundamental idea of the proposed system was to divide PA into different groups based on how they relate to each other in terms of space and meaning, and to build a new framework of hierarchical multi-task CNN for recognizing (global and local) attributes through global sharing and competitions among groups by learning features that can tell them apart. There are both global and local attributes for pedestrians. Most of the time, the global ones depict the whole body features as a whole. While, the local ones are split into several parts: the head and shoulders, the upper-body, and the lower-body. Each of these parts was further split into two groups based on how it looks and what it does. Then, the utilized CNN model is used to measure how well semantic attributes were learned. The framework for multi-task learning lets every CNN model share visual information at the same time with different groups of attribute categories. This framework works better on the PETA dataset, but it requires that the connections between local and global attributes be taken into account explicitly in graphical models.

While Kai Han et al. [8] suggested using the co-occurrence priori to improve CNN performance, especially attribute aware pooling, in 2019. Using a multi-branch CNN as the fundamental architecture, two (co-occurrence) tables were made from the sets of training, and context information from each branch was used for completing the decision made by every branch. This shows not only what different parts of a person are like, but also how they relate to each other. distinct branches. [9] Qiaozhe Li et al. discussed PA recognition as a sequential attribute prediction challenge and proposed a framework of visual-semantic graph logic for addressing this issue. In this approach, there are both spatial and directed semantic graphs. By reasoning with the GCN, one graph records geographical linkages between areas, while the other discovers potential semantic relations among characteristics. A throughout structure is described for performing mutual embedding between these two graphs in order to facilitate relational learning between them. Besides two large-scale datasets, the PETA dataset was utilized to verify the proposed framework. The proposed framework is shown to be more effective in recognizing some hard attributes against a variety of challenging factors. Qiaozhe Li et al. [10] also presented a graph based reasoning module to adaptively bridge visual features and semantic attributes and to perform global reasoning among attribute groups to jointly model their spatial and semantic relations. Additionally, a regularization term was proposed by distilling auxiliary human parsing knowledge to guide the visual-semantic reasoning and enhance feature representations. Experiment results show why the method you want to use is better than the method in [9] and the effectiveness of this reasoning module and auxiliary human parsing knowledge distillation. Haoran An et al. [11] Part-guided Network is a deep network architecture that puts features into six groups based on body parts and uses an attention mechanism to improve convolutional feature maps. To split up the tasks of recognizing PA. The first part of this network is a module for learning backbone features. The second part is a module for part-guided attention. The GoogleNet backbone feature learning module, which is based on the inception architecture, creates representations that are informative. The module for part-guided attention is used to check if the spatial attention mechanism's restrictions on location are true. Different analyses of the PETA dataset showed that the Part-guided Network can accurately find qualities that are very close to where they are located. This shows that the network works. Xingting He et al. [12] came up with a method for grouping attributes-based multi-task convolutional neural networks. This method takes advantage of spatial correlations between attributes and makes sure that each attribute is also somewhat independent. This is different from most previous works, which only dealt with the problem of very unevenly distributed attributes. Also, an online batch weighted loss was made to reduce the differences in performance

between characteristics and improve the average accuracy of the model's recognition. The whole network can be trained from one end to the other, and tests on the PETA dataset show that this is a very effective way to do it.

Yang Li et al. [13] introduced a CNN, channel attention (CAtt), and convolutional Long Short-Term Memory (ConvLSTM) model in 2020. In this model, pre-trained CNN and CAtt are used to get the salient and associated visual properties of pedestrians. Then, using ConvLSTM, more spatial information and correlations are taken from the characteristics of the people walking by. The CAtt mechanism is then paired with ConvLSTM to find out how attributes and the spatial information of their visual features are related. PA are predicted one at a time using a prediction sequence that takes into account the size of the area and how important each attribute is. Using the PETA dataset, experiments were done, and the results effectively took into account the most important attribute characteristics and kept the spatial information of visual features for making predictions about different attributes. To improve the performance of PA recognition even more, it is suggested that the attention mechanism be put in different parts of ConvLSTM so that the semantic and spatial links between PA can be mined more efficiently. Also, this proposed model is locked and doesn't change based on the data set before it. analysis; it is not adaptable to new datasets., thus the prediction sequence should be automatically determined and dynamic based on the connection between attributes. Zhong Ji et al. [14] presented a multiple-time steps attention method that is shown to be successful in the PETA dataset for addressing two important challenges: complex interactions between photos and attributes, and uneven distribution of PA. This technique, unlike existing attention strategies that just focus on the present and past time steps, utilized knowledge about the forthcoming time step. By recording the information of numerous time steps in an adaptable manner, more contextual knowledge is utilized. In order to alleviate the problem of an imbalance in the distribution of pedestrian qualities, a targeted balance loss function was devised by increasing the cost of harder-to-recognize attributes. However, more visual attention processes should be investigated in order to correctly detect the attention areas and record the correlations between visual appearances and qualities with greater precision. Joint Learning of Attribute and Contextual Relations was described by Zichang Tan et al. [15] as an end-to-end network for PA recognition. It is comprised of two unique modules: the attribute relation module and the contextual relation module. For the first module, attribute-specific features are acquired using limited losses, and a Graph Convolutional Network (GCN) is utilized to discover the relationships between several attributes. For the second module, it was recommended to utilize a graph projection approach to project the 2D feature map into a set of nodes from various image areas and then to use GCN to scout the contextual links between these regions. In addition, these two modules are combined into a single framework so that they may be learned simultaneously.

Yang et al. [16] introduced cascaded Split-and-Aggregate learning will be used in 2021 for capturing the uniqueness and similarity for the whole attributes, one at the level of the feature-map and another one at the level of the feature-vector. Concerning the first, utilizing a designing attribute-specified attention model, the features of every attribute were divided. Each attribute's split features are learned with limited losses for the latter. Numerous layers of convolutional or fully-connected are used to combine the separated features of each module. Also, to learn the modules together and at the same time, feature recombination was proposed in this study to execute a random shuffle based on the split features across a batch of data in order to generate additional training examples that span the unevenness of the prospective samples. Jian Jia et al. [17] introduced the consistency framework for PA recognition, which exploits the inter-image relationship with the similar attribute and addresses the spatial attention region divergence issue. Specifically, it was proposed that the spatial consistency module pay attention to certain attribute-related geographical areas. In addition, the Semantic consistency module was proposed for extracting discriminative and substantial semantic features to every attribute. In addition, two-variants of this procedure were utilized to illustrate the efficacy of consistency regularizations. Experiments on ablation shown that both of these consistency modules may provide performance enhancements and consistently reached remarkable PETA performance.

2) Pedestrian Attribute (PA-100K) Dataset

Concerning the fine-grained tasks, learning the inclusive pedestrian features is still an open issue. Therefore, a dataset of large-scale pedestrian attributes called "PA-100K" was introduced by Xihui Liu et al. [18] for facilitating different analysis tasks of pedestrians. PA-100K dataset constructed comprises one hundred thousand pedestrian photos from five hundred and ninety-eight settings, and hence give the most comprehensive dataset for PA recognition. This dataset defines 26 widely used features, including global variables such as gender and age, as well as object-level attributes such as purse, phone, upper clothing, etc. This dataset was assembled using photos recorded by actual outside surveillance cameras, which is more difficult. In contrast to The images were produced by sampling frames from surveillance films, letting future applications like attribute recognition based on video and frame-by-frame evaluation of pedestrian quality work. All of the pedestrians in each photo were labeled, even those whose movements were blurry or whose quality was very low (less than 50100). The whole set of data is randomly separated into 10,000 testing, 10,000 validation, and 80,000 training. Samples from a single person's tracklets in a surveillance video were randomly put into one of these sets. In this case, this dataset makes sure that the learned characteristics are independent of the person's identity. Each of the 26 qualities is shown in both good and bad ways in these sets.

Xihui Liu et al. [18] have suggested Hydra Plus-Net (HPnet) is an attention-based deep neural network that sends information from the multi-level attention mappings in different directions to different feature layers. The Attentive Feature Net is made up of several branches of multidirectional attention (MDA) modules that are applied to different semantic feature levels. The attentive deep features learned from the presented HP-net have some unique benefits. First, the model is capable of capturing multi-attention from low to semantic levels. Second, the model looks into the multiple scale selectivity of features for improving the last representations of features for an image of pedestrian. The HP-net for pedestrian analysis was used to do two things: recognize PA and re-identify people. In general, this indicated that HP-net sought to augment the global feature representation for attribute classification with additional pose, component, or viewpoint data. Nonetheless, our technique has trouble localizing the regions corresponding to distinct properties. Pengze Liu et al. [19] presented a localize guided neural network that allocates weights for specific attributes for locating features depending on the similarity between the locations of the pre-extracted proposals and attributes. In this presented method, the local-features are auto-learned for every attribute and stressed via the interacting with global-features. The efficiency of this presented network was illustrated using PA-100K dataset, and the obtained outcomes surpass the work in [18]. However, these techniques tended to gain attribute-specific characteristics through the building of diverse complicated network topologies with extra modules. Such extra modules significantly increase the amount of network parameters, though. In the meanwhile, the issues of class imbalance and difficult attribute retrieval continue to be undervalued in PA recognition. Lin Chen et al. [20] studied the optimization of the training processing process and presented a function of multiple labels contrastive focal loss. This presented function emphasized minority and hard attributes for mitigating the influence of the non-balance by utilizing a diverse re-weighting scheme for various negative and positive categories. Furthermore, this function was capable of widening the holes between the characteristics of intra-class multiple labels to push CNN for extracting more discriminative nuanced features. Besides the dataset of PA-100K, several big datasets were also utilized in this model. The obtained outcomes outperform the competing related models.

The majority of PA recognition systems highlight a variety of criteria, including visual attention-based, part-based, and attribute-based, etc. While Weichen Chen et al. [21] introduced a view-attribute localization strategy based on attention, which uses information about the view to tell the recognition process to pay attention to certain attributes and the attention mechanism to localize specific attribute-corresponding areas. The view prediction branch uses information about the view to make four view weights that show how confident different features are based on the view. The view weights that come out of this are then used to build certain view-attributes that will help extract deep features and keep an eye on them. Regional attention was added to collect geographical information and record how the view attribute interacts with other channels. This was done so that the location of a view attribute could be looked at. After that, a fine attentive attribute-specific region was found, and regional weights for the view-attribute from different spatial locations were given to the regional attention. Combining view weights with regional weights produces the final outcome of view-attribute recognition. Using the PA-100K dataset, transfer learning was examined to validate the applicability and stability of view prediction and view-attribute supervision.

3) Pedestrian Intention Estimation (PIE) Dataset

PIE dataset created by A. Rasouli et al. [22] included over Six hours of film acquired by a The Waylens Horizon has 157 wide-angle lenses and a calibrated monocular dashboard camera. All videos are captured in HD quality (19201080) at 30 frames per second. Inside the car, the camera was put in behind the rearview mirror. For ease of use, videos are broken up into six sets of about 10 minutes each. The whole set of data was taken in downtown Toronto, Canada, in the afternoon when it was sunny and cloudy. This set of data includes places with a lot of foot traffic and narrow streets as well as places with less foot traffic and wide boulevards. It shows a wide range of pedestrian crossing patterns. The PIE dataset has long, continuous sequences that have been labeled and can be used in many different ways. For each pedestrian close to the road who might interact with the car, the following annotations were added: bounding boxes with occlusion flags, crossing intention confidence, and text labels for pedestrian activities ("walking", "standing", "looking", "not looking", "crossing", "not crossing"). From the time a pedestrian walks into the picture until the time they leave, they can be followed. The data collection has annotations that are needed for perception and visual reasoning, such as bounding boxes for traffic objects, pedestrian intentions and behaviors, and pedestrian attributes (for example, age and gender). There are 1842 samples of pedestrians, which are split into 50 percent train, 40 percent test, and 10 percent validation sets, respectively. A. Rasouli et al. [23] proposed a multitask learning framework for predicting pedestrian trajectory and action. This framework was based on a bifold mechanism for encoding and decoding various input tasks and modalities, thus providing the model to learn cross-correlation between them and inducing it to learn better representations. Furthermore, a novel technique was introduced that implicitly models interactions between target pedestrians and their surroundings by depending on changes in semantic representations of the scenes. Using a publicly available PIE dataset, this proposed method showed significant improvement over existing methods on both trajectory and action prediction tasks with an accuracy of 91%.

Javier Lorenzo et al. [24] proposed a self-attention alternative, based on transformer architecture. This design consists of several branches that combine video and kinematic data. RubiksNet and TimeSformer are the two alternative designs for the video division. The kinematic branch is based on various transformer encoder designs. Several studies with an emphasis on pre-processing input data have been conducted, exposing issues with two kinematic data sources: post key points and ego-vehicle speed.

R. Quan et al. [25] introduced a Holistic LSTM network for pedestrian trajectory prediction, which assessed at every time step the location and movements of the target pedestrian, the speed of the vehicles, the motion dynamics of the global views, and the intentions of the pedestrian to cross the street. People have thought of adding more memory cells, like the speed cell, intention cell, and correlation cell, to LSTMs as a way to make them better at simulating future dynamic fluctuations. Also, a unique gated shifting operation was made to dynamically include pedestrian intention and global correlation information, which largely determined how the pedestrian moved through space. Also, the output of Holistic LSTM was rescaled on the fly to take into account changes in vehicle speed. This made predictions of pedestrian bounding boxes more accurate. In trials, the suggested pedestrian trajectory prediction method achieves a high level of performance utilizing the PIE dataset.

4) Richly Annotated Pedestrian (RAP) Dataset

Dangwei Li et al. [26] built a dataset of RAP produced from actual multi-camera surveillance settings with long-term data collection, where data samples are not only labeled with fine-grained human characteristics, but also environmental and contextual aspects. RAP consisted of 41,585 samples of pedestrians, each of which was annotated with 72 features and perspectives, including 69 binary variables and three multi-value attributes, like shooting angle, occlusion, body part information, and police-provided data. Three consecutive months of footage are captured by 26 cameras with a resolution of 1280 x 720 and a frame rate of 15. RAP includes only data from actual video surveillance scenarios, which are more indicative of actual application settings, in contrast to other datasets that collect data from numerous data sources. Due to perspective, occlusion, etc., the same recognizable individual may have several attribute annotations in the RAP dataset. With PID annotations, however, the same property may behave differently from several viewpoints, therefore identification-based annotation may be incorrect. Considering the cost of annotation, for the final annotation, 17 hours of synchronous videos are chosen by hand from the total number of videos collected. In Table 2, all of the annotations in the RAP dataset are fully explained. There is information about space and time, the whole body, parts, accessories, postures and activities, occlusion, and occlusion. Compared to the pedestrian attribute datasets that are already available, the RAP dataset has four more types of annotations: views, occlusions, human parts, and fine attributes. For more description for the RAP dataset, see Figures 4 and 5.

Table 2: RAP dataset annotations

Classes		**Attribute**
Spatial Temporal		Position of Image, Time, Bounding Box of Body, Accessories, Lower-Body, Upper-Body, and Head-Shoulder.
Whole		Gender, Age, Body Shape, Role.
Accessory		Single-Shoulder Bag, Backpack, Paper Bag, Plastic Bag, Handbag, etcetera.
Posture & Action		Gathering, Telephoning, Carrying, Talking, Viewpoints, Pushing, etcetera.
Occlusion		Occlusion Kinds, Occluded Parts.
Part	LOWER	Footwear Color & Style, Clothes Color & Style.
	UPPER	Clothes Color & Style.
	HEAD	Hair Color & Style, Glasses, and Hat.

Figure 4: Attribute examples in RAP dataset.

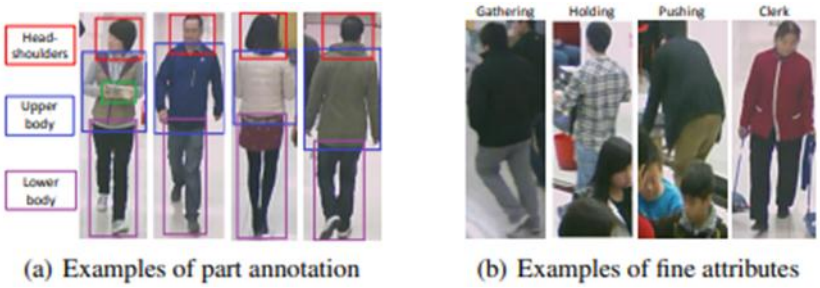


Figure 5: The details of RAP dataset .

Ehsan Yaghoubi et al. [27] A methodology for deep learning was proposed that enhances by removing the background areas from the layers of the network that are fully connected. To do this, a layer of element-wise multiplication was added between the output of the last convolutional layer and a binary mask of the full-body human foreground. Also, the updated feature maps were made smaller and spread out over a number of fully linked layers, each of which is taught to do a certain job (i.e., a subset of attributes). Last but not least, a loss function was made that gives each category of qualities a certain amount of weight to make sure that each attribute is taken into account and that none of the attributes affect the results of the others. The experimental examination of the RAP dataset showed that the proposed model performs much better than the current state of the art. Imran N. Junejo and colleagues [28] used trainable Gabor wavelet layers into a CNN model. The hyper parameters for the Gabor wavelet were created using a neural network, and the resultant Gabor filters were used to filter inputs. The network train consisted of fifty epochs. The model transforms the input picture to grayscale before passing it through a sequence of six mixed-layers blocks that discover the optimal parameters for the output Gabor filters. The efficiency of the presented model was evaluated on the RAP, and the obtained accuracy was 91.1%.
Imran N. Junejo, and Naveed Ahmed [29] proposed a depth wise separable CNN for solving the PA recognition problem. This 24-layer network reduced the number of trainable parameters with almost 50% fewer parameters while making learning efficient. The proposed method was tested on and evaluated using the RAP dataset and obtained recognition accuracy was 97.84% which is a significant improvement over the state of the art.

III. PEDESTRIAN ACTIVITY RECOGNITION

Vision-based pedestrian activity recognition uses cameras to watch how people act and how the environment changes. The techniques of computer vision (CV) like marker extracting, structural modeling, segmentation of motion, extraction of actions, and tracking

motions are used in this technology. Researchers use many different kinds of cameras that combine several cameras for depth or stereo-vision cameras capable of detecting the scene depth with lights of infrared [30].

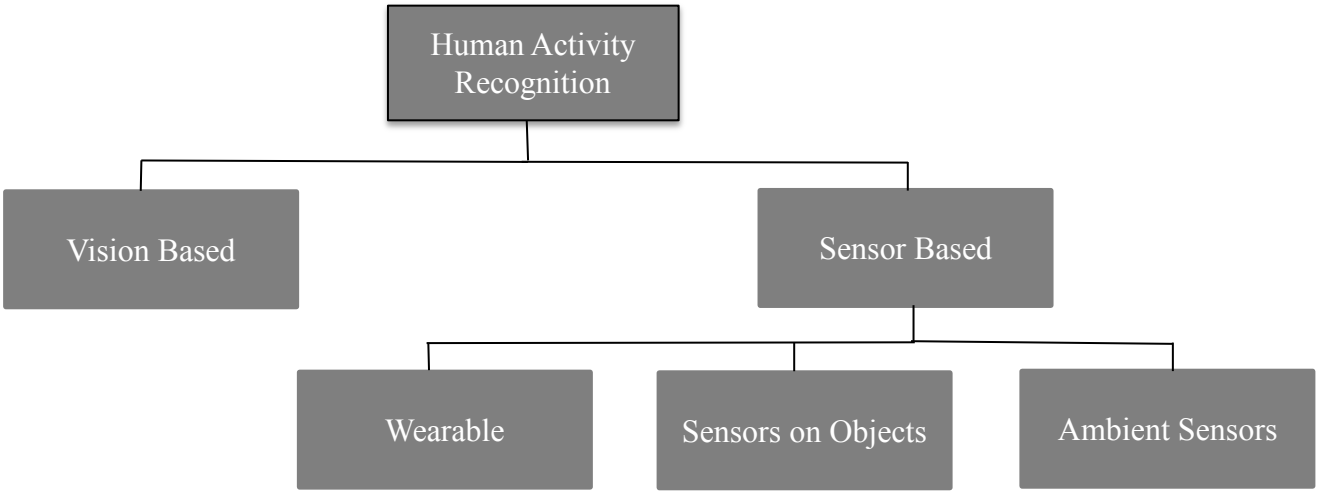


Figure 6: Human activity recognition approaches [30]

Understanding pedestrian Behaviors are a hard topic in CV that has made high progress in recent years. Image and signal processing, feature extraction, machine learning, and three-dimensional geometry are all needed to understand human behavior [31]. Intelligent video systems' main job is to be able to automatically recognize people and understand what they are doing. In the past few years, people's interest in understanding behavior has grown quickly. This is due to social needs like natural interfaces, security, effecient computing, assisted living, and gaming. Hardware and communication protocols have come a long way, which makes it possible for new services to be offered, like collecting data on team sports in real time [13] and annotating films so that events can be found and retrieved [14]. To look at the scene at different levels of abstraction, you have to go through a number of processing steps, starting with how the objects of interest move. First, the subject(s) of interest are found and tracked to create motion descriptions (such as a motion trajectory or a local movement composite), that are analyzed for finding interactions or actions. When evaluating local motions, it's important to be able to tell the difference between intra-body gestures and motion patterns [16]. Depending on how good the camera view is, position information can be supplemented with things like joint trajectories [15] or changes in head attitude [17]. For example, to recognize certain actions, you need to make a set of templates that show different kinds of behaviors. [31].

1) Daimler Dataset

In Daimler Dataset [32], A stereo camera system (22 cm baseline, 16 fps, 1176 640 pixels) positioned behind the windscreen of a car captured picture sequences. Figure 7 depicts four common pedestrian motions: (walking sideways and crossing), (walking sideways and halting), (walking sideways and crossing) (standing at the curb and) initiating (lateral movement), and (walking beside the roadway and bending in and crossing). The collection contains 68 sequences, of which 12485 photographs depict people (singularly). Fifty-five sequences were taken at vehicle speeds between 20 and 30 km/h; the remaining sequences included a vehicle traveling at speeds between 20 and 30 km/h. at rest.

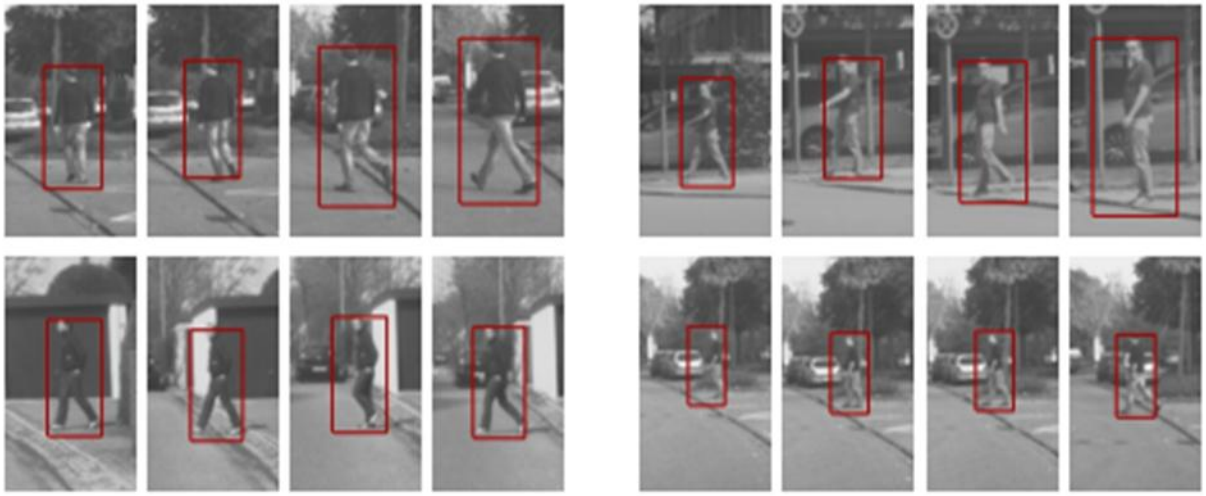


Figure 7: Four normal kinds of pedestrian motions with bounding boxes: (top left) bending in, (bottom left) crossing, (top right) stopping, and (bottom right) starting [32].

According to Daimler Dataset, Peixin Xue et al. [33] A new network design depending on encoder-decoder LSTM was suggested. An encoder of double channels is intended for extracting state streams from the trajectories of pedestrian and vehicle. The decoder then implements state fusion to determine the future trajectory of pedestrians. The experimental findings confirmed the efficacy of this strategy.

Knowing immediately if a detected pedestrian wants to cross the road in front of the vehicle is essential for averting collisions through safe and comfortable navigation. Therefore, Zhijie Fang et al. [34] introduced a novel vision-based scheme that examines the attitude of a pedestrian across many frames for determining if she or he is likely to enter a crosswalk. a traffic region where there is a potential for a collision. Pedestrian detection and skeleton extraction are performed concurrently by using a CNN-based posture estimation approach, which simultaneously recognizes pedestrians and offers them, as well as high-level characteristics collected from their skeletons, to build an efficient classifier. Support Vector Machine (SVM) that processes these features. This method presented experiments demonstrating 750 milliseconds of anticipation for crossing the road, which at a typical urban driving speed of 50 kilometers per hour can provide fifteen additional meters for vehicle automatic reactions or to alert the driver. In addition, unlike previous systems, this method is monocular and does not require stereo or optical flow information. Nonetheless, this strategy necessitates consideration of the identical pedestrian intention situations when more pedestrians are present. obstructing each other, which begins with the production of a suitable dataset including such instances.

2) NTSEL Dataset

Hirokatsu Kataoka et al. [35] gathered 100 movies containing four actions in traffic situations through an experiment. These behaviors involve walking, turning, and bicycle riding, and crossing, all of which demonstrate pedestrian mobility with three individuals. The dataset features a complex backdrop in tiny locations; as a result, it is challenging to detect flows. Presented activities were likewise fine-grained, since there was little variance between walking, crossing, and turning, which occur at almost identical pedestrian movements. 5-fold cross validation was used to analyze this dataset. Changes in pedestrian behavior (such as "straight walking" to "turning") have only fine-grained feature distinctions. Figure 8 illustrates the distinction between pedestrian activities. Evidently, the feature descriptions from the same individual within the self-collected dataset tend to be pretty similar. Particularly, the three actions of "turning", " straight walking", and "crossing" are gaining fine-grained recognition due to their comparable form, changing pedestrian size, and inclusion in the same category as "walking."

Figure 8: Various activities of pedestrians using the collected dataset: (a) Bicycle riding, (b) Turning, (c) Straight walking, and (d) Crossing.[35]

H. Kataoka et al. [36] proposed a basic way for illustrating a change in a flow image using "acceleration images." These pictures should be significant since their representation differs from that of location (RGB) and velocity (flow) images. In this approach, two-stream CNN was used as the starting point, followed by acceleration, spatial, and temporal streams. From a series of flow photos, several computations are performed to generate acceleration images. Automatic feature learning with CNN may considerably extract a crucial feature from acceleration photos, despite the sparse representation's tendency to contain noisy data. The experiments were conducted on traffic data in the NTSEL dataset, and the proper recognition of the acceleration, spatial, and temporal streams are 82.5%, 87.5%, and 77.5%, respectively. While, H. Kataoka et al. [37] proposed a motion descriptor for facilitating transitional actions recognition to specify the sensitive variations between transitional and action actions. In this work, a CNN-based motion descriptor was utilized to successfully present a clear distinction between transitional and action actions. In order to understand pedestrian actions, three datasets including the NTSEL dataset were utilized. The proposed CNN-based provided the best rate of success on three trial datasets. Even when using the shortest (3-frame) feature accumulation for recognition tuning, good results were confirmed with 85.78% on the NTSEL dataset.

3) Collective Activity Dataset (CAD)

In order to classify pedestrian activities, Wongun Choi and S. Savarese [38] created the Collective Activity Dataset (CAD) which was acquired under non-constrained real-world conditions. In the CAD dataset, more than forty short video clips of talking, walking, queueing, waiting, and crossing action classes were registered. The videos had pixels of size 640x480 and were registered using a camera of consumer handheld. Figure 9 demonstrates the scenes' complexity in which the top three rows illustrate proper categorization, whereas the bottom row illustrates improper classification. The estimated horizon is superimposed as a red dashed line on the photos. When predicted poses deviate greatly from real postures or when collective activity is not well-defined by people's movement, the categorization method fails (4th crossing example).

Figure 9: Example results.

Each 10th frames of all video sequences was manually tagged with ground truth information, including posture, activity, and bounding box data. For learning reasons, only the posture label is needed. The residual labels are utilized to evaluate performance and characterize the dataset. The key properties of the CAD dataset are detailed in Table 3.

Table 3: The main characteristics of the CAD.

Properties	Waiting	Talking	Crossing	Walking	Queueing	Overall
No. of People	4.19	3.86	3.89	2.57	7.32	5.22
No. of Classes	1.39	1.49	1.42	1.46	1.15	1.37
Activity Clutter	0.31	0.49	0.35	0.41	0.14	0.33
Bounding Box Overlap	0.48	0.38	0.24	0.16	0.43	0.34
Camera Shake	13.76	12.88	18.55	19.53	23.30	18.35

Zhiwei Deng et al. [39] presented A hierarchical graphical model based on deep neural networks for recognizing individual and group activities in surveillance situations. As an initial phase, deep neural networks are employed to identify the behaviors of specific individuals in a scene. Then, a hierarchical graphical model based on neural networks refines the projected labels for each activity by incorporating interclass relationships. Similar to the inference mechanism in a probabilistic graphical model, the refining phase imitates the message-passing incorporated into the architecture of a deep neural network. Using the CAD dataset, our method demonstrated effective group activity recognition with an accuracy of 80.6%. Z. Deng et al. [40] they came up with a way to combine graphical models and deep neural networks into one framework. This end-to-end trainable deep network classified low-level image inputs according to their content and refined these classifications by passing messages between outputs. Furthermore, it provides structure learning by gating functions that specify which outputs to connect and resulted in an efficient categorization with 81.2% of accuracy.

Ibrahim et al. [41] designed a deep model for capturing individual people's activity (temporal dynamics) based on LSTM models. A LSTM model is made to show how people's actions change over time, and another LSTM model is made to combine information about each person to understand the whole activity. The CAD dataset was used to test this model. Experimental results show that the proposed model improves group activity recognition performance with an accuracy of 81.5%. A hierarchical attention and context modeling framework for group activity recognition was presented by L. Kong et al. [42]. Hierarchical Context Networks model intra-group and inter-group contextual information. Hierarchical Attention Networks devote varying amounts of attention to various individuals and their diverse body regions. The suggested approach can create more discriminating descriptions of group activities by incorporating visual attention and contextual structure. Extensive studies on the CAD dataset reveal that the suggested framework is superior with an accuracy of 84.3 percent.

Zalluhoglu and Ikizler-Cinbis [43] introduced a multi-stream architecture for collective activity recognition based on person regions. In addition to employing global picture data, this system assessed several local locations for determining collective actions. This system is the first to recognize collective action using a network of multi-stream and different locations. In addition, it evaluated several techniques for combining multi-spatial and temporal streams to increase the recognition accuracy. The experimental evaluation of the CAD dataset achieved an accuracy of 88.9 percent, representing a significant increase.

IV. PERFORMANCE EVALUATION AND ANALYSIS

In order to evaluate the algorithms of pedestrian attribute and activity recognition, the mean accuracy (M_{Ac}) can be utilized. In every attribute or activity, M_{Ac} computes the categorization accuracy of negative and positive samples and obtains their values of average as the attribute or activity recognition result. Lastly, the rate of recognition can be gotten via getting average overall attributes or activities. This metric of evaluation is computed using the following equation [44]:

$$M_{Ac} = \frac{1}{M}\sum_{i=1}^{L}\left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}\right) \quad (1)$$

Where TNi and TPi indicate the number of rightly predicted negative and positive examples, Ni and Pi indicate the number of negative and positive examples, respectively, and L indicates the number of attributes. The above-mentioned metric of evaluation handles every attribute separately and disregards the correlation of the inter-attribute that prevails in the multiple attribute recognition issues. This metric is named as a label-based criterion and proposes to utilize the example-based evaluation metric motivated via the fact that this type of evaluation catches superior the consistency of prediction on the provided image of pedestrians.
The commonly utilized metrics of evaluation (F1 metric (F), accuracy (Ac), precision (P), and recall (R)) are given as follows [45]:

$$Accuracy = \frac{1}{M}\sum_{i=1}^{L}\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|} \quad (2)$$

$$Recall = \frac{1}{M}\sum_{i=1}^{L}\frac{|Y_i \cap f(x_i)|}{|Y_i|} \quad (3)$$

$$Precision = \frac{1}{M}\sum_{i=1}^{L}\frac{|Y_i \cap f(x_i)|}{|f(x_i)|} \quad (4)$$

$$F_1 = \left(\frac{2\times Precision\times Recall}{Precision+Recall}\right) \quad (5)$$

Where Y_i indicates the ground-truth positive labels of the example (i), $f(x)$ proceeds the positive labels of predicted for example (i), $|\cdot|$ indicates the set cardinality, and M indicates the number of examples.

Table 4: Comparison among pedestrian attributes and activity recognition datasets.

Datasets	Recognition	Classes

PETA	Attributes	Age16-30, Age31-45, Age46-60, AgeAbove61, Male, Backpack, messenger bag, Plastic bag, Carrying-Others, No carrying, Muffler, Hat, Formal upper, Formal lower, Logo, Casual upper, Casual lower, Shoes, Leather shoes, Sandals, Sneaker, Trousers, Jeans, Skirt, Shorts, Jacket, V-Neck, Stripes, Short-Sleeve, Plaid, T-shirt, Upper-Other, Long hair, Sunglasses, No accessory
PA-100K	Attributes & Activities	Age Less16, Age 17-30, Age 31-45, Clerk, Customer, Female, Body Normal, Body Fat, Body Thin, Bald Head, Black Hair, Long hair, Bag, Paper Bag, Hand-Bag, Plastic Bag, Backpack, Hand Trunk, Box, Dress, Suit-Up, Vest, Jacket, Shirt, T-shirt, Sweater, Short Sleeve, Long Trousers, Tight Trousers, Jeans, Skirt, Short Skirt, Casual Shoes, Cloth Shoes, Cotton, leather shoes, Sports Shoes, Boots, Hat, Glasses, Other Attachment, Muffler, Tight, Pusing, Carrying by Hand, Carrying by Arm, Pulling, Talking, Gathering, Holding, Calling
PIE	Attributes	gender, age
	Activities	Standing, Walking, Not-Looking, Looking, Not-Crossing, Crossing.
RAP	Attributes & Activities	Age Less16, Age 17-30, Age 31-45, Female, Customer, Clerk, Body Fat, Body Thin, Body Normal, Black hair, Bald head, Box, Handtrunk, Backpack, Handbag, Hat, Dress, Jacket, Jeans, Cloth shoes, Casual shoes, Cotton, Boots, Glasses, Gathering, Carrying by hand, Carrying by arm, Holding, Calling, etcetera.
Daimler	Activities	crossing, stopping, bending in, starting
NTSEL	Activities	walking, crossing, turning, riding a bicycle
CAD	Activities	crossing, waiting, queueing, walking, talking

TABLE 5: COMPARISON AMONG PEDESTRIAN ATTRIBUTES AND ACTIVITY RECOGNITION METHODS USING DEEP LEARNING.

Ref.	Authors & year	RAP					PETA					PA-100k				
		M_{Ac}	Ac	P	R	F	M_{Ac}	Ac	P	R	F	M_{Ac}	Ac	P	R	F
[1]	Junejo (2021)	---	91.5	92.59	91.6	91.9	---	80.1	84.77	80.1	81.79	---	---	---	---	---
[7]	Fang. et al. (2018)	83.25	63.13	82.52	81.65	82.08	88.20	78.31	86.23	89.21	87.69	---	---	---	---	---
[8]	Han1. et al. (2019)	81.42	68.37	81.04	80.27	80.65	86.97	79.95	87.58	87.73	87.65	80.56	78.30	89.49	84.36	86.85
[9]	Li. et al. (2019)	77.91	70.04	82.05	80.64	81.34	85.21	81.82	88.43	88.42	88.42	79.52	80.58	89.40	87.15	88.26
[10]	Li. et al (2019)	78.30	69.79	82.13	80.35	81.23	84.90	80.95	88.37	87.47	87.91	77.87	78.49	88.42	86.08	87.24
[11]	An. et al(2019)	84.63	60.85	74.72	74.49	74.60	95.18	94.27	95.83	96.81	96.32	---	---	---	---	---
[12]	He. et al. (2019)	81.43	67.95	78.46	81.46	79.93	85.73	79.88	87.39	86.79	87.09	---	---	---	---	---
[13]	Li. et al. (2020)	83.72	---	81.85	79.96	80.89	88.56	---	88.32	89.62	88.97	---	---	---	---	---
[14]	Ji. et al. (2020)	77.62	67.17	79.72	78.44	79.07	84.62	78.80	85.67	86.42	86.04	---	---	---	---	---
[15]	Tan. et al. (2020)	83.69	69.15	79.31	82.40	80.82	86.96	80.38	87.81	87.09	87.45	82.31	79.47	87.45	87.77	87.61
[16]	Yang. et al. (2021)	84.18	68.59	77.56	83.81	80.56	86.40	79.93	87.03	87.33	87.18	82.86	79.64	86.81	88.78	87.79
[17]	Jia. et al. (2021)	82.77	68.37	75.05	87.49	80.43	86.52	78.95	86.02	87.12	86.99	81.87	78.89	85.98	89.10	86.87
[18]	Liu. et al. (2017)	76.12	65.39	77.33	78.79	78.05	81.77	76.13	84.92	83.24	84.07	74.21	72.19	82.97	82.09	82.53
[19]	Liu. et al. (2018)	78.68	68.00	80.36	79.82	80.09	---	---	---	---	---	76.96	75.55	86.99	83.17	85.04
[20]	Zheng. et al. (2021)	82.06	69.01	77.47	84.91	81.02	86.84	78.78	83.68	89.97	86.71	81.11	79.01	86.67	88.15	87.41
[21]	Chen. et al. (2022)	78.33	67.48	79.81	80.84	80.32	---	---	---	---	---	80.08	78.14	87.60	86.73	87.16
[28]	Junejo. et al. (2021)	---	91.1	92.39	91.1	91.56	---	80.04	86.49	80.1	82.32	---	---	---	---	---
[29]	Junejo and Ahmed (2021)	91.33	97.84	78.56	66.60	72.07	91.08	80.06	82.96	77.88	80.32	---	---	---	---	---

V. CONCLUSION

Pedestrian attributes and activity recognition represents significant research concentrating on image detection and categorization field. In this comprehensive survey, we presented the recent pedestrian attributes and activity recognition methods using deep learning. Particularly, the issue formulation and challenges of pedestrian attributes and activity recognition were introduced. The most utilized benchmarking datasets were also introduced as well as the proposals that are using them. Furthermore, we provided a comparison using the most common metrics for evaluating the recognition methods However, owing to the restricted space in this survey, several related methods and the main construction of the deep learning algorithms are not covered. Therefore, these missed concepts will be summarized in future works.

REFERENCES

[1] I. N. Junejo, "Pedestrian attribute recognition using two-branch trainable Gabor wavelets network," PLoS One, vol. 16, no. 6, 2021.
[2] Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

[3] T. M. Hasan, S. D. Mohammed, J. Waleed, "Development of Breast Cancer Diagnosis System Based on Fuzzy Logic and Probabilistic Neural Network", Eastern-European Journal of Enterprise Technologies, Information and Controlling System, Vol. 4, No. 9 (106), pp. 6-13, 2020.
[4] Deng, Y.; Luo, P.; Loy, C.C.; Tang, X. Pedestrian attribute recognition at far distance. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 789–792.
[5] Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, Bin Luo,Pedestrian attribute recognition: A survey, Pattern Recognition, Vol. 121,2022.
[6] Chen, X., Zhuang, S., Zheng, X., & Wang, Z. (2021, December). Pedestrian Attribute Recognition Based On Deep Learning: A Survey. In *2021 International Conference on Information Technology and Biomedical Engineering (ICITBE)* (pp. 140-144). IEEE.
[7] Fang W., Chen J., Lu T., Hu R. (2018) Pedestrian Attributes Recognition in Surveillance Scenarios with Hierarchical Multi-task CNN Models. Advances in Multimedia Information Processing – PCM 2018. PCM 2018. Lecture Notes in Computer Science, vol 11165. Springer, Cham.
[8] Han, K., Wang, Y., Shu, H., Liu, C., Xu, C., & Xu, C. (2019). Attribute aware pooling for pedestrian attribute recognition. arXiv preprint arXiv:1907.11837.
[9] Qiaozhe Li, Xin Zhao, Ran He, Kaiqi Huang. Visual-semantic graph reasoning for pedestrian attribute recognition. In: Proceedings of the AAAI conference on artificial intelligence. 2019. p. 8634-8641.
[10] Qiaozhe Li, Xin Zhao, Ran He, Kaiqi Huang. "Pedestrian Attribute Recognition by Joint Visual-semantic Reasoning and Knowledge Distillation." IJCAI. 2019.
[11] H. An, H. Fan, K. Deng and H. -M. Hu, "Part-guided Network for Pedestrian Attribute Recognition," 2019 IEEE Visual Communications and Image Processing (VCIP), 2019, pp. 1-4,
[12] X. He, Q. Shi, F. Su, Z. Zhao and B. Zhuang, "Pedestrian Attribute Recognition Based on Mtcnn with Online Batch Weighted Loss," 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 2461-2465.
[13] Li, Y.; Xu, H.; Bian, M.; Xiao, J. Attention Based CNN-ConvLSTM for Pedestrian Attribute Recognition. Sensors 2020, 20, 811
[14] Zhong Ji, Zhenfei Hu, Erlu He, Jungong Han,Yanwei Pang,Pedestrian attribute recognition based on multiple time steps attention, Pattern Recognition Letters, Volume 138,2020,Pages 170-176,
[15] Tan, Z., Yang, Y., Wan, J., Guo, G., & Li, S. Z. (2020, April). Relation-aware pedestrian attribute recognition with graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 12055-12062).
[16] Yang, Y., Tan, Z., Tiwari, P. et al. Cascaded Split-and-Aggregate Learning with Feature Recombination for Pedestrian Attribute Recognition. International Journal of Computer Vision, 129, 2731–2744, 2021.
[17] Jia, J., Chen, X., & Huang, K. (2021). Spatial and Semantic Consistency Regularizations for Pedestrian Attribute Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 962-971).
[18] Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 350–359.
[19] Liu, P., Liu, X., Yan, J., & Shao, J. (2018). Localization guided learning for pedestrian attribute recognition. arXiv preprint arXiv:1808.09102.
[20] X. Zheng, Z. Yu, L. Chen, F. Zhu and S. Wang, "Multi-label Contrastive Focal Loss for Pedestrian Attribute Recognition," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 7349-7356.
[21] Chen, WC., Yu, XY. & Ou, LL. Pedestrian Attribute Recognition in Video Surveillance Scenarios Based on View-attribute Attention Localization. Mach. Intell. Res. (2022).

[22] A. Rasouli, I. Kotseruba, T. Kunic and J. Tsotsos, "PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6261-6270
[23] A. Rasouli, M. Rohani and J. Luo, "Bifold and Semantic Reasoning for Pedestrian Behavior Prediction," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15580-15590..
[24] Javier Lorenzo, Ignacio P. Alonso, Rubén Izquierdo, Augusto L. Ballardini, Álvaro H. Saz, David F. Llorca, and Miguel Á. Sotelo, "CAPformer: Pedestrian Crossing Action Prediction Using Transformer", Sensors, Vol. 21, No. 17: 5694, 2021.
[25] R. Quan, L. Zhu, Y. Wu and Y. Yang, "Holistic LSTM for Pedestrian Trajectory Prediction," in IEEE Transactions on Image Processing, vol. 30, pp. 3229-3239, 2021.
[26] Li, D., Zhang, Z., Chen, X., Ling, H., & Huang, K. (2016). A richly annotated dataset for pedestrian attribute recognition. arXiv preprint arXiv:1603.07054..
[27] Ehsan Yaghoubi, Diana Borza, João Neves, Aruna Kumar, Hugo Proença, "An attention-based deep learning model for multiple pedestrian attributes recognition", Image and Vision Computing, Vol. 102,2020.
[28] Imran N. Junejo, Naveed Ahmed, Mohammad Lataifeh, "Pedestrian attribute recognition using trainable Gabor wavelets", Heliyon, Vol. 7, No. 6, 2021.
[29] Imran N. Junejo, and Naveed Ahmed, "Depthwise Separable Convolutional Neural Networks for Pedestrian Attribute Recognition", SN Computer Science, Vol. 2, no. 100, 2021.
[30] Bouchabou, D.; Nguyen, S.M.; Lohr, C.; LeDuc, B.; Kanellos, I. A Survey of Human Activity Recognition in Smart Homes Based on IoT Sensors Algorithms: Taxonomies, Challenges, and Opportunities with Deep Learning. Sensors 2021, 21, 6037.
[31] P. V. K. Borges, N. Conci and A. Cavallaro, "Video-Based Human Behavior Understanding: A Survey," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 11, pp. 1993-2008, Nov. 2013.
[32] Schneider N., Gavrila D.M. (2013), Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study. Pattern Recognition. GCPR 2013. Lecture Notes in Computer Science, vol 8142. Springer, Berlin, Heidelberg
[33] P. Xue, J. Liu, S. Chen, Z. Zhou, Y. Huo and N. Zheng, "Crossing-Road Pedestrian Trajectory Prediction via Encoder-Decoder LSTM," 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019, pp. 2027-2033.
[34] Fang, Z., Vázquez, D., & López, A. M. (2017). On-board detection of pedestrian intentions. Sensors, 17(10), 2193.
[35] H. Kataoka, Y. Aoki, Y. Satoh, S. Oikawa and Y. Matsui, "Fine-Grained Walking Activity Recognition via Driving Recorder Dataset," 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015, pp. 620-625.
[36] H. Kataoka, Y. He, S. Shirakabe, Y. Satoh, "Motion Representation with Acceleration Images". Computer Vision, ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, Vol. 9915. Springer, Cham, 2016.
[37] H. Kataoka, Y. Miyashita, M. Hayashi, K. Iwata, Y. Satoh, "Recognition of Transitional Action for Short-Term Action Prediction using Discriminative Temporal CNN Feature", Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, pp. 1-12, 2016.
[38] Wongun Choi and S. Savarese, "A unified framework for multitarget tracking and collective activity recognition', In Computer Vision–ECCV 2012, pages 215–230. Springer, 2012.
[39] Zhiwei Deng, Mengyao Zhai, Lei Chen, Yuhao Liu, Srikanth Muralidharan, Mehrsan Javan Roshtkhari, Greg Mori, "Deep structured models for group activity recognition." arXiv preprint arXiv:1506.04191 (2015).
[40] Z. Deng, A. Vahdat, H. Hu and G. Mori, "Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4772-4781.
[41] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat and G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1971-1980.
[42] L. Kong, J. Qin, D. Huang, Y. Wang and L. Van Gool, "Hierarchical Attention and Context Modeling for Group Activity Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1328-1332.
[43] C. Zalluhoglu, N. Ikizler-Cinbis, "Region based multi-stream convolutional neural networks for collective activity recognition", Journal of Visual Communication and Image Representation, Vol. 60, pp. 170-179, 2019.
[44] Waleed, J., Albawi, S., Flayyih, H. Q., & Alkhayyat, A. (2021, September). An Effective and Accurate CNN Model for Detecting Tomato Leaves Diseases. In 2021 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA) (pp. 33-37). IEEE.
[45] Waleed, J., Abbas, T., & Hasan, T. M. (2022, March). Facemask Wearing Detection Based on Deep CNN to Control COVID-19 Transmission. In 2022 Muthanna International Conference on Engineering Science and Technology (MICEST) (pp. 158-161). IEEE.