A Survey on People Re-identification Approaches Considering Occlusion

Zahra Mortezaie[1*], Hamid Hassanpour[2]
1- Department of Computer Engineering, Shahrood University of Technology, Shahrood, Iran.
Email: zm.mortezie@gmail.com (Corresponding author)
2- Department of Computer Engineering, Shahrood University of Technology, Shahrood, Iran.
Email: h.hassanpour@shahroodut.ac.ir

Received: April 2022 Revised: May 2022 Accepted: July 2022

ABSTRACT:
Analyzing human behavior and detecting visual anomaly are important applications of video surveillance systems in many fields such as security systems and intelligent buildings. Person re-identification (RE-ID) is one of the main steps in a surveillance system, where, it has directly an effect on system performance. Occluded body parts, backgrounds clutter, and variations in pose and scene illumination are some noticeable problems in appearance-based RE-ID approaches, as they have an effect on pedestrians' appearance during tracking task. Among the mentioned problems, person RE-ID considering appearance changes caused by occlusion can be considered as a common problem in video surveillance systems. In this paper, some existing people RE-ID approaches are briefly reviewed in terms of robustness to person body occlusion. Also, the experimental results reported in these approaches are compared using some partial occluded databases. The comparison results demonstrate the supremacy of the non-pose-guided RE-ID approaches.

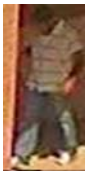
KEYWORDS: Person Re-Identification, Surveillance System, Occlusion, Pose-guided.

1. INTRODUCTION

Video surveillance systems consist of a network of cameras with overlapping or non-overlapping field of view, in order to monitor the position and behavior of people in public and private places such as airports [1], train stations [2], shopping centers [3], and intelligent buildings [4]. The environment covered by these cameras can be the interior or exterior of a building, or a combination of both. One of the important points in video surveillance systems is re-identifying persons accurately [5]. Person RE-ID is used to determine labels of persons in a camera's field of view and re-identifying them after a while, whether in the same camera's field of view or in the field of view of another cameras [6]. Person RE-ID as one of the most effective steps in surveillance systems, is a necessary prerequisite for high level processing such as analyzing the persons behavior and recognizing events [7]. Hence, accurate person RE-ID improves high level processing, and consequently improves the performance of video surveillance systems.

Re-identifying persons in each of the internal and external environments has its own problems. Partial occluded body parts, backgrounds clutter, and variations in pose and scene illumination are some noticeable problems in RE-ID approaches as they lead variations in persons' appearance [8]. Note that, variations in persons' appearance reduces the accuracy of appearance-based RE-ID approaches.

In Fig. 1, the mentioned problems are shown using some images from the DukeMTMC-reID [9], VIPeR [10], GRID [11], and 3DPeS [12] databases. In each column of the figure, a pedestrian at two different camera views is shown. In sample 1, the person body is occluded by the car. Also, in sample 3, the carried object occludes body parts. In sample 2, in the field of view of different cameras, the background of the same person changes. Note that, in holistic characteristic extraction approaches, the occluded regions as well as background are treated as the persons' appearance. Meanwhile, according to samples 3 and 4, variations in person pose and scene illumination across different cameras affect the visual appearance characteristics.

Sample 1	Sample 2	Sample 3	Sample 4
			

Paper type: Research paper
DOI: https:// 10.30486/mjee.2022.696501
How to cite this paper: Z. Mortezaie, H. Hassanpour, **"A Survey on People Re-identification Approaches Considering Occlusion"**, *Majlesi Journal of Electrical Engineering*, Vol. 16, No. 3, pp. 9-18, 2022.

Fig. 1. Some images from the DukeMTMC-reID, VIPeR, GRID, and 3DPeS databases.

Considering the above-mentioned points, in this paper we briefly review and compare the performance of some people RE-ID approaches in terms of robustness to human body occlusion. In this field, some RE-ID approaches are proposed based on several auxiliary modules such as pose estimation and semantic segmentation approaches, whereas another RE-ID approaches are introduced without depending on auxiliary modules. Hence, in this paper we categorize the RE-ID approaches into two groups as: the RE-ID approaches which are based on the guidance of pose estimation approaches; and those that do not use the guidance of pose estimators.

This paper is organized as follows. The occluded person RE-ID approaches are briefly reviewed in Section 2. The commonly used databases for occluded people RE-ID are introduced in Section 3. The performance comparison between the reviewed approaches is depicted in Section 4. Finally, the conclusions are driven in Section 5.

2. PEOPLE RE-ID APPROACHES CONSIDERING OCCLUSION

As mentioned in Section 1, the occluded pedestrian RE-ID approaches are categorized into two groups as: the RE-ID approaches that are guided using pose estimation approaches, namely pose-guided RE-ID approaches; and those that do not use the guidance of pose estimators, namely non-pose-guided RE-ID approaches. In this section, we briefly review some RE-ID approaches that are in the mentioned groups.

2.1. Pose-guided RE-ID approaches

Wang et al., [13], proposed a pedestrian RE-ID framework for learning high-order relation and topology information of the person body. In this framework, semantic local characteristics are extracted from a Convolutional Neural Network (CNN) based on the key points estimated by a human key points estimator. Then, by considering the local characteristics as nodes of a graph, an Adaptive Direction Graph Convolutional (ADGC) layer is proposed for passing relation information between nodes. Since, aligning two groups of local characteristics from two images is considered as a graph matching problem, in [13] a Cross-Graph Embedded Alignment (CGEA) layer is used to learn topology information and jointly embed topology information into local characteristics.

In [14], the occluded and non-occluded body parts are determined by training a part label generator. In this generator, a pose estimator is applied to extract the key points of the person body, where, it estimates the coordinates and confidence score of each key point. Considering the images as horizontal regions, the number of key points in each region and their confidence scores are used in a voting mechanism to decide about the existence of occlusion in the region. In the voting mechanism, the confidence score of a key point is treated as a weight which is used in voting whether the corresponding region is occluded or not. The higher confidence score denotes the higher voting weight. In this mechanism, the existence of occlusion in each region is judged considering a threshold for the summation of the corresponding key points' weights. Finally, in order to determine the visibility scores of body parts without needing key points, a discriminator is tuned using the characteristics of each horizontal region extracted from a CNN and the part labels generated by the generator in an end-to-end manner.

The RE-ID approach introduced by Miao et al.,[15], consists of three branches based on the estimated human body poses. In the first branch, namely Pose-Masked Characteristic Branch, a pre-trained pose estimator extracts the key points of the person body. Considering a threshold on the confidence scores of the key points, the occluded and non-occluded body parts are estimated. The non-occluded parts are then used to provide spatial masks in order to guide the CNN in focusing on the informative regions of the images (i.e., non-occluded body parts). In the second branch, namely Pose-Embedded Characteristic Branch, a pose embedding is generated using the non-occluded body parts. Using this branch as the gates in the CNN, leads to adaptively tune the response of the network channels, where, the gates active the channels associated with non-occluded parts, as well as, suppress the channels related to the occluded body parts. In the third branch, for determining the similarity between probe and gallery images, first, the non-occluded regions of the images are determined using the estimated pose. Then, the local characteristics correspond to the non-occluded body parts are used for comparing between the images.

Xu et al., [16], tried to detect the occluded and non-occluded regions of the person body, and decrease the effects of occluded region on extracted characteristics and matching steps. To achieve this goal, in this approach, a pedestrian RE-ID framework, namely dual-attention RE-ID (DAReID), is proposed, where it involves two branches as a mask branch and a global branch. In the mask branch, a Pose-Guided Spatial Attention approach (PGSA) is used in order to obtain the occluded and non-occluded body parts. In this branch,

for occluded parts, pose guided coarse labels are obtained using key points of the person body. The obtained coarse labels are then used to guide the back-bone network (i.e., ResNet-50 [17]) to learn local characteristics. Besides, in the global branch, the visual activation levels of different regions are obtained using activation-based attention approach. The obtained visual activation levels are merged with pedestrian pose information in order to define Weighted Local Distances (WLD). In addition, the WLD learning approach is used for forcing the back-bone network to learn more discriminative local characteristics.

Zheng et al., [18], proposed a people RE-ID framework, namely Pose-Guided Characteristic Learning with Knowledge Distillation [19] (PGFL-KD). This framework includes a Main Branch (MB), and two pose-guided branches, namely a Foreground Enhanced Branch (FEB), and a body part Semantics Aligned Branch (SAB). In FEB, the foreground characteristic alignment is performed by emphasizing the characteristics of non-occluded body parts, as well as, ignoring the interference of occluded regions and background. In SAB, for obtaining body part semantics aligned representation, different channel groups are forced for focusing on different body parts. Also, MB is trained using both the characteristics extracted in FEB and SAB in a knowledge distillation and interaction-based training manner.

Ma et al., [20], proposed a Pose-guided Inter-part and Intra-part Relational Transformer (PIRT) approach, assuming fixed positions of a pedestrian in the images. In this approach, first, a pose-guided characteristic extraction module is used for extracting key points, and then expanding their corresponding regions into person body masks. In this approach, the expanded regions are merged into three groups as upper, middle, and lower parts of body in order to extract local characteristics from body parts. Hence, the generated masks and extracted characteristics from a back-bone network are used as the input of Intra-part Relational Module (IRM), where, they guide IRM to create the local relations in each group. In IRM, a Global Average Pooling (GAP) layer is utilized for merging the information in each group as local characteristics. These characteristics are then passed throughout the Inter-part Relational Transform (IRT) module, where, it obtains cross relationships between parts as the structural characteristics using transformers. For re-identifying the pedestrians, the nearest neighbor of the probe image is obtained based on extracted local and structural characteristics.

Wang et al., [21], proposed a framework, namely Shrinking and Re-weighting Network (SRNet). SRNet learns global characteristics via shrinking as well as re-weights local characteristics simultaneously. In this framework, ResNet50 is used as a back-bone network, where, it is improved by adding the Deep Residual Shrinkage (DRS) module [22], to the stages 1 to 3 of the back-bone for eliminating different noisy characteristics caused by occlusion. To achieve this goal, DRS module learns the soft thresholds for converting the near-zero characteristics to zeros. The extracted global characteristics from the improved back-bone are then converted to the local semantic part characteristics, using the heat maps of the key-points obtained from a pose estimator, namely HRNet [23]. The extracted global and local characteristics are then passed throughout the Re-weight Module for Part Matching (RMPM) in order to learn self-adaptive weights for global and local characteristics in Re-weight Module (RM). Using the RM, the effects of characteristics extracted from the occluded body parts are reduced in part matching.

2.2. Non-pose-guided RE-ID Approaches

Yan et al., [24], proposed a RE-ID model to extract discriminative single-scale global-level characteristics from the non-occluded body parts. In this model, first, occluded samples are generated using a data augmentation approach, namely Compound Batch Erasing (CBE). Then, the occluded and non-occluded samples are passed throughout the ResNet-50 in order to optimize a proposed Bounded Exponential Distance (BED) loss function. In BDE lose function, an exponential penalization enables the model to learn various local discriminative parts in pairs of images with high similarities. Also, in this RE-ID model, the reconstructive pooling layer of the back-bone (i.e., ResNet-50), is modified using Disentangled Non-Local operation (DNL) added to the second/third stage of ResNet-50 for forcing the network to focus on the non-occluded body parts.

In [25], the occluded pedestrian RE-ID is formulated as a set matching problem. In this approach, a CNN is used as a back-bone network to extract characteristics from images. The extracted characteristic vector of an image is passed throughout a non-linear activation function and the output of the activation function is considered as a pattern set. Each pattern set is represented using a global vector, where, each element of the global vector denotes a specific visual pattern. The similarity between the images is determined based on their global vectors using a proposed Jaccard similarity coefficients as the metric. In the proposed metric, minimization and maximization operations are used to respectively simulate the intersection and union operations used in classic Jaccard similarity in order to provide the proposed metric applying on continuous real numbers, as well as, training network. Also, in order to use the set matching problem in training CNN, a Jaccard triplet loss is proposed in [25].

Jia et al., [26], proposed a RE-ID approach based on

disentangled representation learning, namely DRL-Net. DRL-Net involves a CNN back-bone and encoder-decoder layers based on the transformer architecture [27], used for extracting compact representations, as well as, generating characteristics of semantic component. Using transformer architecture in DRL-Net and accordingly global reasoning of local characteristics existed in occluded person images bring alignment free RE-ID framework. In this approach, using learnable positional encodings, spatial information is encoded into label-relevant characteristics for matching between images, and label-irrelevant characteristics for occlusion interference elimination. The encoded information is added to the input of the encoder attention layers. Also, a set of learnable input is defined for decoder layers, as semantic preferences object queries to obtain semantic components characteristics. Using the objects queries the representations of undefined semantic components existed in the occluded person images are disentangled. Besides, decoder of the transformer imposes a decorrelation constraint on semantic preference object queries in order to make them to focus on related semantic components.

Li et al., [28], introduced a Part-Aware Transformer (PAT) approach for people RE-ID under occlusion. Similar to [26], this approach is based on transformer architecture. PAT consists of a transformer encoder and a transformer decoder. In the transformer encoder, a self-attention approach is used to extract context information from the image. Also, in order to obtain context aware characteristic map robust to background clutters, the correlation of the characteristic map pixels is modeled, where, similar pixels are aggregated. In the decoder, a set of learnable part prototypes is defined for generating part-aware masks, where, they focus on discriminative person body parts. In this approach, the part-aware masks are obtained by determining the similarity between the characteristic map pixels and part prototypes, where, a part-aware mask contains the spatial distribution of a specific body part. Using these masks, characteristics of person body parts are obtained by a weighted pooling. Also, part diversity and part discriminability approaches are introduced in [28] for guiding the part prototype learning. The part diversity approach is used for obtaining lower correlation between characteristics of part and forcing part prototypes to focus on different discriminative foreground regions. Besides, the part discriminability approach is used for forcing characteristics of part to identity discriminative using part classification and a triplet loss. Finally, the transformer encoder and decoder are simultaneously optimized in order to learn part prototypes.

Hou et al, [29], attempted to recover the occluded body parts, where, they proposed a characteristics completion block, namely Region Characteristic Completion (RFC), for video-based occluded person RE-ID. In this approach, an Adaptive Partition Unit (APU) is proposed for dividing the characteristic maps extracted from the back-bone network adaptively into different regions, where, each region is related to a specific body part. Then, based on a region-encoder and a region decoder, a Spatial Region Characteristic Completion (SRFC) module is proposed for recovering the occluded regions using spatial information. In this module, first, the region-encoder learns the correlation between non-occluded and occluded body parts. Then, the occluded body parts and corresponding non-occluded parts are aggregated to an intermediate node using the region encoder. Besides, the spatial correlation is used in the region-decoder for recovering occluded region characteristics. In this step, the characteristics of the occluded body parts are estimated from the intermediate node using the region-decoder. Indeed, in SRFC for recovering characteristics of occluded parts, the information from corresponding non-occluded body parts is propagated to the occluded parts via the intermediate nodes.

Wang et al., [30], proposed a Self-guided Body Part Alignment approach (SBPA), where, it obtains significant body parts in order to extract discriminative characteristics. This approach, first, generates the scale-wise Global Attention Modules (GAM) with cross-scale information, where, they are used for automatically determining the significant body parts. In this approach, Local Relation Transformer (LRT) networks are used for independently predicting semantic-aligned local parts. Using relation message passing via a transformer network, the predicted semantic-aligned parts are represented in this step. Also, the local predictions are guided to be semantic using a Self-guided Constraint Loss (SCL), where, it minimizes the difference between local and anchored global attentions. Meanwhile, the similarity between the extracted characteristics from the images are computed and further are merged based on a calculated score visibility of the body parts in order to reduce the effects of the occluded parts.

Jin et al., [31], introduced a supervised people segmentation module, namely attribute-guided occlusion-sensitive pedestrian segmentation (AOPS). AOPS is trained using randomly occluded sample images in order to segment the occluded images. Then, shift characteristic adaption (SFA) module uses the body part masks obtained from AOPS for generating the discriminative appearance characteristics. In SFA, first, the non-occluded samples are passed throughout the occlusion batch painting (OBP) module in order to generate the occluded samples. Then, the segmentation masks and characteristics of generated occluded samples are used as the input of a shift attention representation (SAR) module in order to extract the shifted visible characteristics from the samples. The extracted characteristics and corresponding masks are passed

throughout a mask-based drop block (MDB) module for refining the extracted characteristics. Finally, a visible region matching (VRM) approach is used in test step, for eliminating useless parts of non-occluded person body. In this approach, the segmentation masks of the gallery images are substituted with the mask of the probe image. Using the mask of the probe image, SFA is guided to extract characteristics from those regions of the gallery images which are visible in the probe image.

Chen et al., [32], proposed an Occlusion-Aware Mask Network (OAMN), where, it involves three parts as: an occlusion augmentation module, an attention-guided mask module, and an occlusion unification module. The occlusion augmentation module generates various occluded sample images by adding randomly scaled rectangular patches to the images. In attention-guided mask module, spatial weight maps are generated for each input characteristic via RGA-S [33]. This module is trained based on the occluded samples and spatial weight maps to extract characteristics from person body parts in occluded and non-occluded images. For re-identifying probe image, the occlusion unification module is used, where, it is trained using the occluded samples in order to learn a supervised grader. The learned grader, determines the location of occlusion existed in the images considering four locations (i.e., top, bottom, left, right), and two areas (i.e., half, quarter). After obtaining the location of Occluded body parts in the probe image, the same location in the gallery images is occluded in order to reduce ambiguity of the probe image.

Huang et al., [34], proposed an adversarial model, where, it contains two CNNs (CNNs) and a discriminator. The CNNs are used to learn characteristics of non-occluded and occluded images, and the discriminator is used for distinguishing between them. In this approach, for generating the occluded samples, first, the salient regions of the attention maps are considered as the location of Occluded body parts. Then, the corresponding location in the non-occluded image is substituted with a constant random value. The occluded sample generation is used simultaneously in each iteration of the CNN training process in order to learn the location of Occluded body parts, as well as, generate various training occluded samples. The occluded images are then used for confusing a discriminator in distinguishing between the characteristics of the original non-occluded and occluded samples. In re-identifying the probe images, both the characteristics of the original non-occluded images and occluded image are integrated for representing the pedestrian images, in order to improve the accuracy of matching between the images.

Zhang et al., [35], proposed an attention mechanism, where, they guide the back-bone CNN to focus on the non-occluded body parts. The main idea of the proposed attention mechanism, is that different channels of CNN-based pedestrian detectors (in this research i.e., FasterRCNN pedestrian detector [36]) are selective and show responses for different parts of body. Hence, in [35], a channel-wise attention mechanism is added to the characteristic extractor network (i.e., ResNet50), in order to generate attention parameters (i.e., weights), for different channels of the extracted convolutional characteristics, where, lower weights are assigned to the channels correspond to the occluded regions, whereas, higher weights are assigned to the channels correspond to the body parts. Consequently, in this approach, using the attention mechanism in the back-bone network forces the network to focus on the characteristics of the body parts.

3. COMMON DATABASES USED FOR OCCLUDED PERSON RE-ID

In this section, we briefly review some commonly used databases for assessing the performance of occluded person RE-ID approaches, i.e., Occluded-DukeMTMC [37], Occluded-ReID [38], Partial-REID [39], Partial-iLIDS [40], Market1501 [41], and DukeMTMC-ReID [9].

Partial-REID database involves 600 images from 60 pedestrians captured from a university campus with various viewpoints and occlusions. For each person, it has 5 non-occluded (holistic) and 5 occluded (partial) images in gallery and query sets, respectively. In Fig. 2, some data samples from Partial-REID database are shown. Each column of this figure is related to a data sample.

Fig. 2. Some data samples from Partial-REID database.

Partial-iLIDS database is collected based on iLIDS database [42], where, it involves 238 images from 119 pedestrians captured via multiple non-overlapping cameras in the airport. Also, in this database the

occluded regions of the samples are manually cropped. In Fig. 3, some data samples from Partial-iLIDS database are shown. Each column of this figure is related to a data sample.

Fig. 3. Some data samples from Partial-iLIDS database.

Note that, as the Partial-REID and Partial-iLIDS are small databases, the existing occluded person RE-ID approaches usually use these databases in the test stage.

Occluded-REID is another occluded person image database, where, it involves 2000 images from 200 occluded persons. This database has 5 occluded images as well as 5 non-occluded images for each pedestrian, where, they are captured using mobile cameras. In Fig. 4, some data samples from Occluded-REID database are shown.

Fig. 4. Some data samples from Occluded-REID database.

Market-1501 database involves a few occluded samples. Hence, it is usually used for assessing non-occluded person RE-ID approaches. Meanwhile, this database is used to assess the generalization ability of the occluded person RE-ID approaches. Market-1501 involves 32688 images from 1501 pedestrians captured via 6 cameras. The training, gallery, and query sets of this database have 12936, 19732, and 3368 images respectively. In Fig. 5, some data samples from Market-1501 database are shown.

Fig. 5. Some data samples from Market-1501 database.

DukeMTMC-reID database involves 1404 pedestrians captured via 8 cameras, where, it has 16522, 2228, and 17661 images in training, query, and gallery sets respectively. Similar to Market-1501, this database is used to assess the generalization ability of the occluded person RE-ID approaches.

Occluded-DukeMTMC is the largest database used for occluded person RE-ID. It involves the occluded images of DukeMTMC-reID database, where, it has 15618, 17661, and 2210 images in training, gallery, and query sets respectively. Note that in this database, all of the query images are occluded. In Fig. 6, some data samples from DukeMTMC-reID database are shown. Each column of this figure is related to a data sample.

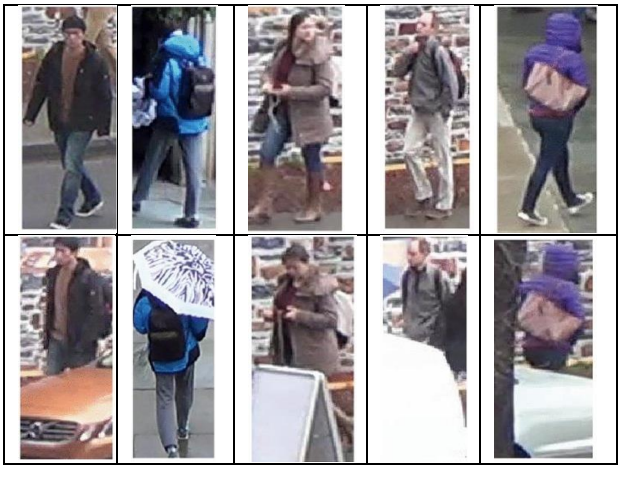


Fig. 6. Some data samples from DukeMTMC-reID database.

4. PERFORMANCE COMPARISON BETWEEN OCCLUDED PERSON RE-ID APPROACHES

In this section, the experimental results from some existing occluded person RE-ID approaches are reviewed based on the Occluded-DukeMTMC, Occluded-ReID, Partial-REID, Partial-iLIDS, Market1501, and DukeMTMC-ReID databases. The performance of these approaches is compared using the reported Rank-1. Rank-k (k = 1, 2, 3, …) is a common measure used to assess the performance of people RE-ID approaches. The accuracy of RE-ID approaches is usually treated as a function of Rank-k, where, for

ranking the obtained results from a RE-ID approach, k denotes that k-top matches the correct answer [43, 44]. Consequently, Rank-k is a strictest measure for k = 1, whereas, this measure permits some error for k > 1. In Table 1, the performance of some occluded person RE-ID methods is compared based on the best reported results in Rank-1.

Table 1. Comparison between the performance of occluded person RE-ID approaches based on Rank-1.

		Occluded databases				Non-occluded databases	
	Approaches	Occluded-DukeMTMC	Occluded-ReID	Partial-REID	Partial-iLIDS	Market1501	DukeMTMC-ReID
Pose-guided	Wang et al., [13]	55.1	80.3	85.3	72.6	94.2	86.9
	Yang et al., [14]	62.2	81.0	85.7	80.7	-	-
	Miao et al., [15]	56.3	-	72.5	70.6	92.7	86.2
	Xu et al., [16]	63.4	-	76.7	68.1	94.6	88.9
	Zheng et al., [18]	63.0	80.7	85.1	74.0	95.3	89.6
	Ma et al., [20]	60.0	-	-	-	94.1	88.9
	Wang et al., [21]	65.5	80.6	86.0	-	94.5	87.3
Non-pose-guided	Yan et al., [24]	69.0	78.5	-	-	96.1	91.1
	Jia et al., [25]	66.6	-	-	-	-	-
	Jia et al., [26]	65.0	-	-	-	94.7	88.1
	Li et al., [28]	64.5	81.6	88.0	76.5	95.4	88.8
	Hou et al., [29]	63.9	-	-	-	95.2	90.7
	Wang et al., [30]	64.5	-	-	-	96.0	89.6
	Jin et al., [31]	55.4	82.5	86.8	81.7	94.6	87.5
	Chen et al., [32]	62.6	-	86.0	77.3	93.2	86.3
	Huang et al., [34]	-	-	-	-	95.5	87.9
	Zhang et al., [35]	-	-	53.7	46.2	92.8	81.1

Wang et al., [13], focused on the high-order relation and topology information of the person body. In this approach, persons' key points are estimated by a pose estimation approach. Then, semantic local characteristics were extracted from a CNN based on the estimated key points. Also, Yang et al., [14], used a voting mechanism for determining the visibility scores of body parts, where, it was based on the human key points. In this mechanism, the images are considered as horizontal regions and, in each region, the existence of occlusion was determined based on the number of key points as well as their confidence scores. In [15], the occluded and non-occluded body parts were estimated considering a threshold on the confidence scores of the key points extracted by a pre-trained pose estimator. In this approach, the estimated non-occluded parts were used to provide spatial masks for guiding the CNN in order to focus on the non-occluded body parts. Xu et al., [16], used a pose-guided spatial attention approach for determining the occluded and non-occluded body parts. In this approach, for occluded parts, pose guided coarse labels were obtained via key points of the person body. Then, the obtained coarse labels were used for guiding the CNN to learn local characteristics. Besides, the RE-ID module proposed in [18], involves two pose-guided branches, namely FEB and SAB branches. Meanwhile, a pose-guided characteristic extraction module was used in [20], to extract human key points, as well as, expand their corresponding regions into person body masks, where, it was used in order to extract local characteristics from body parts. In [21], the global appearance characteristics were converted to the local semantic part characteristics, using the heat maps of the key-points obtained from a pose estimator. Not that, at least one stage of the above-mentioned RE-ID approaches, is based on the results obtained from a pose estimator module. However, the pose estimation approaches suffer from some errors in estimating the human key points. Consequently, the performance of mentioned pose-guided RE-ID approaches directly depends on the performance of pose estimators.

Yan et al., [24], used a data augmentation approach, namely CBE, to create occluded samples from non-occluded images. Then ResNet-50 was trained using the occluded and non-occluded samples in order to learn various local discriminative parts in pairs of images with high similarities. Meanwhile, in this approach, the reconstructive pooling layer of the ResNet-50, was modified by adding a DNL operation to the second/third stage of the network. This modification along with training the network using both the occluded and non-occluded samples, force the network to automatically focus on the non-occluded body parts. Hence, according to the reported results in Table 1, the re-dentification approach proposed in [24], outperforms other comparing approaches on Occluded-DukeMTMC, DukeMTMC, and Market1501 databases. Besides, in [25], first, a CNN was used to extract characteristics from images. Then, the extracted characteristic vector was converted to a pattern set using a non-linear activation function. The pattern set was represented using a global vector, where, each element of the global vector denoted a specific visual pattern. Besides, the similarity between the images was considered as matching between the sets. Hence, Jaccard similarity coefficients as the metric was

used in [25] to determine similarity between the global vectors. Converting the row extracted characteristics to the specific visual patterns, improves the performance of this approach on Occluded-DukeMTMC comparing other non-pose-guided RE-ID approaches (excepted [24]).

Jia et al., [26], provided alignment free RE-ID framework, using transformer architecture and accordingly global reasoning of local characteristics existed in occluded person images. In this approach, spatial information was encoded into label-relevant characteristics and label-irrelevant characteristics. The encoded information was added to the input of the encoder attention layers. Also, a set of learnable input was defined for decoder layers, as semantic preferences object queries to obtain semantic components characteristics. The representations of undefined semantic components in the occluded samples were disentangled using the objects queries. Similar to [26], the RE-ID approach proposed in [28], was based on transformer architecture, where, it consisted of a pixel context-based transformer encoder and a part prototype-based transformer decoder. In the pixel context-based transformer encoder, a self-attention approach was used to extract context information from throughout the image. In the part prototype-based decoder, a set of learnable part prototypes was defined to generate part-aware masks, where, the masks were used to focus on discriminative person body parts. Besides, in [29] the characteristic maps extracted from the back-bone network adaptively were divided into different regions, where, each region was related to a body part. Then, to recover the occluded regions, SRFC module was proposed based on a region-encoder and a region decoder. In this module, the region-encoder learns the correlation between non-occluded and occluded body parts. The spatial correlations were used in the region-decoder for recovering occluded region characteristics. In this approach, the characteristics of the occluded body parts were estimated from the intermediate node using the region-decoder. Also, in [30], LRT networks were used to independently predict body parts. Similar to the pose-guided RE-ID approaches, relying on the obtained results from encoder and decoder layers, imposes some errors in these RE-ID approaches.

The RE-ID approaches proposed by Jin et al., [31], was based on a supervised people segmentation module, namely AOPS, where, it was trained using randomly occluded sample images. The body part masks obtained from AOPS were then used to extract the discriminative appearance characteristics. In this approach, the segmentation masks of the gallery images were substituted with the mask of the probe image, in order to extract characteristics from those regions of the gallery images which were visible in the probe image. Chen et al., [32], proposed an occlusion unification approach, where, it was trained using the occluded samples to learn a supervised grader. The learned graded, then, determined the location of occlusion existed in the images. After obtaining the location of occluded body parts in the probe image, the same location in the gallery images was occluded for reducing ambiguity of the probe image. Note that ignoring some regions of the gallery images, may lead to miss some discriminative characteristics existed in these regions. Besides, an adversarial model was proposed by Huang et al., [34], where, it contained two CNNs and a discriminator. The CNNs were used to learn characteristics of non-occluded and occluded samples. Meanwhile the discriminator was used to distinguish between the extracted characteristics. In this RE-ID approach, the salient regions of the attention maps were considered as the location of occluded body parts. Hence, the corresponding location in the non-occluded image was substituted with a constant random value in order to generate occluded samples. Note that, the salient regions of the attention maps may denote saliant characteristics in the pedestrians' appearance. Hence, occluding these regions may lead to miss some discriminative appearance characteristics. Meanwhile, in [35], a channel-wise attention mechanism was added to the ResNet50, in order to generate attention parameters (i.e., weights), for different channels of the extracted convolutional characteristics. However, tuning the effects of different channels of the extracted characteristics based on the attention maps, may lead to miss some discriminative characteristics.

Considering the above-mentioned points, developing complex module for occluded person RE-ID imposes some errors according to the errors of the module's parts. This point is especially obvious in some pose guided RE-ID approaches, where, the performance of the RE-ID approaches directly has an effect on the accuracy of the pose estimators. Hence, it is important to develop RE-ID approaches with incorporating minimum auxiliary modules. Besides, all the above-mentioned RE-ID approaches tried to handle the problem of appearance changes caused by occlusion. However, the type of the considered occluded body parts was the occlusion caused by crowded background. Note that, in addition to the occlusion caused by crowded background, the occlusion may be occurred by carried objects. In this situation, the above-mentioned approaches cannot reduce the effects of occlusion on appearance characteristics. Hence, it is necessary to develop RE-ID approaches considering both type of occlusions.

5. CONCLUSION

Pedestrian RE-ID approaches are applied in video surveillance systems, where, the performance of these approaches have directly an effect on the performance of

surveillance systems. This step is used to determine labels of pedestrians in images usually using their visual appearance. It is a challenging task as the appearance may change across the cameras network. Occluded body parts, backgrounds clutter, and variations in pose and scene illumination are some noticeable problems in people RE-ID, as they lead appearance changes. Appearance changes caused by occlusion is one of the important problems in RE-ID approaches. There are some RE-ID approaches considering appearance changes caused by occlusion. In this paper, some occluded person RE-ID approaches are reviewed briefly. Also, the experimental results of these approaches are compared based on a number of partial occluded databases. The experimental results show that the non-pose-guided RE-ID approaches have better performance compared with the other RE-ID approaches.

REFERENCES

[1] T. Van Phat, S. Alam, N. Lilith, P.N. Tran, and N.T. Binh, **"Deep4air: A novel deep learning framework for airport airside surveillance"**, In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1-6, 2021.

[2] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, **"A review of video surveillance systems",** *Journal of Visual Communication and Image Representation*, Vol. 77, p.103116, 2021.

[3] A. Alshammari, and D.B. Rawat, **"Intelligent multi-camera video surveillance system for smart city applications"**, In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0317-0323), 2019, IEEE.

[4] J. Vanus, J. Machac, R. Martinek, P. Bilik, J. Zidek, J. Nedoma, and M. Fajkus, **"The design of an indirect method for the human presence monitoring in the intelligent building"**, *Human-centric Computing and Information Sciences*, Vol. 8, No. 1, pp.1-44, 2018.

[5] G.K. Nayak, U. Shreemali, R.V. Babu, and A. Chakraborty, **"Efficient person re-identification in videos using sequence lazy greedy determinantal point process (slgdpp)",** In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4569-4573, 2019, IEEE.

[6] Q. Leng, M. Ye, and Q. Tian, **"A survey of open-world person re-identification**", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30, No. 4, pp.1092-1108, 2019

[7] Gowsikhaa, D., Abirami, S. and Baskaran, R., (2014) 'Automated human behavior analysis from surveillance videos: a survey', *Artificial Intelligence Review*, Vol. 42, No. 4, pp.747-765.

[8] J. Liu, Z.J. Zha, Q.I. Tian, D. Liu, T.Yao, Q. Ling, and T. Mei, **"Multi-scale triplet cnn for person re-identification",** In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 192-196, 2016.

[9] Z. Zheng, L. Zheng, and Y. Yang, **"Unlabeled samples generated by gan improve the person re-identification baseline in vitro",** In *Proceedings of the IEEE international conference on computer vision*, pp. 3754-3762, 2017.

[10] D. Gray, and H. Tao, **"Viewpoint invariant pedestrian recognition with an ensemble of localized characteristics",** In *European conference on computer vision*, pp. 262-275, 2008, Springer, Berlin, Heidelberg.

[11] C.C. Loy, C. Liu, and S. Gong, **"Person re-identification by manifold ranking"**, In *2013 IEEE International Conference on Image Processing*, pp. 3567-3571, 2013, IEEE.

[12] D. Baltieri, R. Vezzani, and R. Cucchiara, **"3dpes: 3d people database for surveillance and forensics"**, In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pp. 59-64, 2011.

[13] G.A. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, **"High-order information matters: Learning relation and topology for occluded person re-identification",** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6449-6458, 2020.

[14] J. Yang, J. Zhang, F. Yu, X. Jiang, M. Zhang, X. Sun, Y.C. Chen, and W.S. Zheng, **"Learning To Know Where To See: A Visibility-Aware Approach for Occluded Person Re-Identification**", In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11885-11894, 2021.

[15] J. Miao, T. Wu, and Y. Yang, **"Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification",** *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[16] Y. Xu, L. Zhao, and F. Qin, **"Dual attention-based method for occluded person re-identification"**, *Knowledge-Based Systems*, Vol. 212, p.106554, 2021.

[17] K. He, X. Zhang, S. Ren, and J. Sun, **"Deep residual learning for image recognition",** In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[18] K. Zheng, C. Lan, W. Zeng, J. Liu, Z. Zhang, and Z.J. Zha, **"Pose-Guided Characteristic Learning with Knowledge Distillation for Occluded Person Re-Identification"**, In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4537-4545, 2021.

[19] J. Gou, B> Yu, S.J. Maybank, and D. Tao, **"Knowledge distillation: A survey",** *International Journal of Computer Vision*, Vol. 129, No. 6, pp.1789-1819, 2021.

[20] Z. Ma, Y. Zhao, and J. Li, "**Pose-guided Inter-and Intra-part Relational Transformer for Occluded Person Re-Identification"**, In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1487-1496, 2021.

[21] H. Wang, X. Chen, and C. Liu, **"Pose-guided part matching network via shrinking and reweighting for occluded person re-identification",** *Image and Vision Computing*, Vol. 111, p.104186, 2021.

[22] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, **"Deep residual shrinkage networks for fault diagnosis"**, *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 7, pp.4681-4690, 2019.

[23] K. Sun, B. Xiao, D. Liu, D. and J. Wang**, "Deep high-resolution representation learning for human pose estimation"**, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693-5703, 2019.

[24] C. Yan, G. Pang, J. Jiao, X. Bai, X. Feng, and C. Shen, **"Occluded person re-identification with single-scale global representations",** In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11875-11884, 2021.

[25] M. Jia, X. Cheng, Y. Zhai, S. Lu, S. Ma, Y. Tian, and J. Zhang, "**Matching on sets: Conquer occluded person re-identification without alignment",** In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 2, pp. 1673-1681, 2021.

[26] M. Jia, X. Cheng, S. Lu, and J. Zhang, **"Learning Disentangled Representation Implicitly via Transformer for Occluded Person Re-Identification"**, *IEEE Transactions on Multimedia*, 2022.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, **"Attention is all you need",** In *Advances in neural information processing systems*, pp. 5998-6008, 2017.

[28] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, **"Diverse Part Discovery: Occluded Person Re-Identification with Part-Aware Transformer",** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2898-2907, 2021.

[29] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, **"Characteristic Completion for Occluded Person Re-Identification**", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[30] G. Wang, X. Chen, J. Gao, X. Zhou, and S. Ge, **"Self-Guided Body Part Alignment With Relation Transformers for Occluded Person Re-Identification"**, *IEEE Signal Processing Letters*, Vol. 28, pp.1155-1159, 2021.

[31] H. Jin, S. Lai, and X. Qian, **"Occlusion-sensitive person re-identification via attribute-based shift attention",** *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[32] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q.A. Chen, and R. Ji, **"Occlude them all: Occlusion-aware attention network for occluded person re-id",** In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11833-11842, 2021.

[33] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, **"Relation-aware global attention for person re-identification"**, In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 3186-3195, 2020.

[34] W. Huang, S. Liu, R. Luo, T. Si, and Z. Zhang, **"Dynamically occluded samples via adversarial learning for person re-identification in sensor networks",** *Ad Hoc Networks*, Vol. 110, p.102316, 2021.

[35] S. Zhang, D. Chen, J. Yang, and B. Schiele, **"Guided Attention in CNNs for Occluded Pedestrian Detection and Re-identification**", *International Journal of Computer Vision*, Vol. 129, No. 6, pp.1875-1892, 2021.

[36] S. Ren, K. He, R. Girshick, and J. Sun, **"Faster r-cnn: Towards real-time object detection with region proposal networks"**, *Advances in neural information processing systems*, Vol. 28, 2015.

[37] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, **"Pose-guided characteristic alignment for occluded person re-identification",** In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 542-551, 2019.

[38] J. Zhuo, Z. Chen, J. Lai, and G. Wang, **"Occluded person re-identification"**, In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, 2018, IEEE.

[39] W.S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, **"Partial person re-identification",** In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4678-4686, 2015.

[40] W.S. Zheng, S. Gong, and T. Xiang, **"Person re-identification by probabilistic relative distance comparison",** In *CVPR 2011*, pp. 649-656, 2011, IEEE.

[41] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, **"Scalable person re-identification: A benchmark",** In *Proceedings of the IEEE international conference on computer vision*, pp. 1116-1124, 2015.

[42] L. He, J. Liang, H. Li, and Z. Sun, "**Deep spatial characteristic reconstruction for partial person re-identification: Alignment-free approach",** In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7073-7082, 2018.

[43] Z. Mortezaie, and H. Hassanpour, **"A Survey ON AGE-INVARIANT FACE RECOGNITION METHODS"**, *Jordanian Journal of Computers and Information Technology (JJCIT)*, Vol. 5, No. 2, pp.87-96, 2019.

[44] Z. Mortezaie, H. Hassanpour, and A. Beghdadi, **"A Color-Based Re-Ranking Process for People Re-Identification",** In *2021 9th European Workshop on Visual Information Processing (EUVIP)* (pp. 1-5), 2021, IEEE.