International Journal of Network Dynamics and Intelligence Scilight

Survey/review study

A Review of Techniques on Gait-Based Person Re-Identification

Babak Rahi [1,*], Maozhen Li [1], and Man Qi [2]

[1] Department of Electronics and Computer Engineering, Brunel University London, Uxbridge, Middlesex, UB8 3PH, United Kingdom
[2] The School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom
[*] Correspondence: babak.h.rahi@hotmail.com

Received: 16 October 2022
Accepted: 14 December 2022
Published: 27 March 2023

Abstract: Person re-identification at a distance across multiple non-overlapping cameras has been an active research area for years. In the past ten years, short-term Person re-identification techniques have made great strides in accuracy using only appearance features in limited environments. However, massive intra-class variations and inter-class confusion limit their ability to be used in practical applications. Moreover, appearance consistency can only be assumed in a short time span from one camera to the other. Since the holistic appearance will change drastically over days and weeks, the technique, as mentioned above, will be ineffective. Practical applications usually require a long-term solution in which the subject's appearance and clothing might have changed after the elapse of a significant period. Facing these problems, soft biometric features such as Gait has stirred much interest in the past years. Nevertheless, even Gait can vary with illness, ageing and emotional states, walking surfaces, shoe types, clothes types, carried objects (by the subject) and even environment clutters. Therefore, Gait is considered as a temporal cue that could provide biometric motion information. On the other hand, the shape of the human body could be viewed as a spatial signal which can produce valuable information. So extracting discriminative features from both spatial and temporal domains would benefit this research. This article examines the main approaches used in gait analysis for re-identification over the past decade. We identify several relevant dimensions of the problem and provide a taxonomic analysis of current research. We conclude by reviewing the performance levels achievable with current technology and providing a perspective on the most challenging and promising research directions.

Keywords: gait feature extraction; convolutional neural networks; gait re-identification; gait-recognition; neural networks

1. Introduction

The concern over the safety and security of people is continuously growing in recent years. It is no secret that governments are severely concerned with the security of public places such as metro stations, shopping malls, and airports. Protecting the public is an expensive and taxing endeavour. Consequently, governments seek the help of private companies and scientists to alleviate this pressure and provide better security solutions. With the rise of the COVID-19 pandemic, the need for better security solutions is even more evident. Video surveillance systems and CCTV cameras are crucial in optimising such efforts.

The abundance of security cameras and surveillance systems in public areas is valuable for tackling various security issues, including crime prevention. All the recordings from these video surveillance systems must be analysed by surveillance operators, which can be daunting. These operators need to analyse these surveillance videos in real-time and identify various categories of anomalies while looking for a "person of interest" who could easily change their appearance and be unidentifiable to a human with naked eyes. Intelligent video surveillance systems (IVSS) automate the process of analysing and monitoring hours of acquired videos and help the operators understand and handle them. Person re-identification (Re-ID) is one of the most challenging problems in IVSS, which uses computer vision and machine learning techniques to achieve automation.

In the world of computer vision and surveillance systems, person re-identification refers to recognising a person

of interest at different locations using multiple non-overlapping cameras. In other words, identify an individual over a massive network of video surveillance systems with non-overlapping fields of view [1,2]. Person re-identification can also be defined as matching individuals with samples in publicly available datasets, which may have various positions, view angles, lighting or background conditions. This process can be performed on an image sequence or video frames that have been prerecorded or in real-time.

Person Re-ID problems arise when the subject moves from one camera view to another in a network of cameras since moving to another camera view could change their position as well as lighting and background conditions. Even the distance and the angle of the camera view, along with numerous other factors, such as unidentified objects and areas from one camera view to another, can affect the outcome of Re-ID.

Furthermore, person Re-ID is used in security and forensics applications to help the authorities and government agencies find a person of interest. There are three general steps to every person's re-identification solution. Segmentation determines which parts of the frames need to be segmented and focused on. The signature generation finds invariant signatures to compare these parts, and finally, comparison finds an appropriate method to compare these signatures.

There are multiple methods for person Re-ID, but an image or sequence of the subjects is usually given as a Query or Probe. The individuals recorded by the camera network as a template gallery are given as frames. Descriptors are generated for both the template gallery and the probe and consequently compared, where the system gives a ranked list or percentage of similarity based on the probability and similarity of images or sequences to the Probe.

Figure 1 shows a very primitive person Re-ID pipeline in which the system tries to find the corresponding images for a given probe in a gallery of templates. Creating the gallery relies directly on how the re-identification solution has been set up, and [3] categorises this as single-shot and multiple shots, which indicates one or more than one template per frame, respectively. In the case of multiple shots, the new image will be used in the continuous person Re-ID solution, and each time the new image will be used as a probe for the next level.

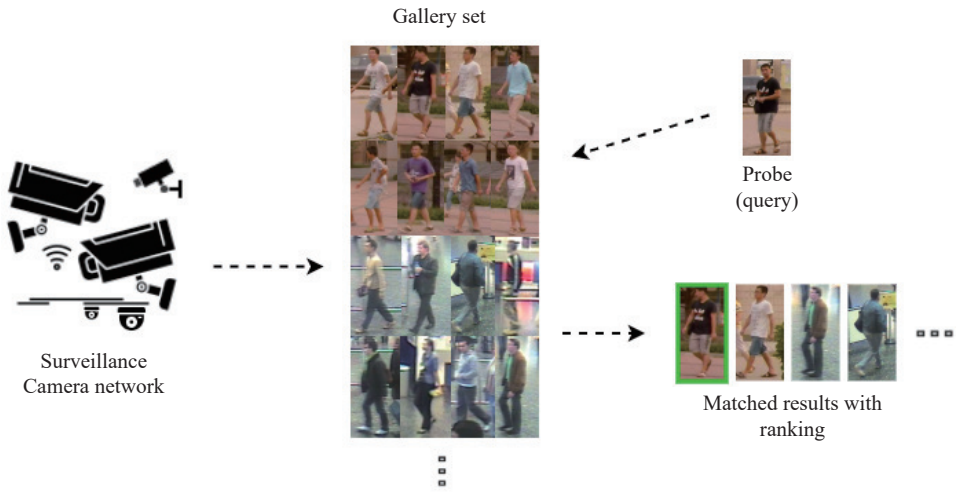

Figure 1. A Basic person Re-ID framework.

Deep learning methods, in particular, convolutional neural networks (CNNs), are used in person Re-ID solutions to learn from a large number of data [4–6]. A CNN uses many filters to look at an image through a smaller window and create feature maps at each layer. Features maps are essentially the reported results of findings after the image has been run through the filter. A particular combination of low-level features can be an indication of more complex features, and feature maps can gradually capture higher-level features at each layer of the network [7].

There are various ways of training a deep neural network (DNN) model. Depending on the problem and the availability of adequate labelled data, these approaches can be categorised as supervised learning, semi-supervised learning and unsupervised learning. In the problem of person Re-ID, mainly for security and anti-terrorism applications, only a small amount of training data is available, so semi-supervised or unsupervised models might result in inaccurate solutions. Another categorisation of DNN person Re-ID approaches is based on their learning methodologies.

These methods can be also be categorised into short-term and long-term person re-identification according to the time interval. Figure 2 shows an example of a short-term person Re-ID in which camera-1 and camera-2 monitor the same walking path with non-overlapping surveillance views.

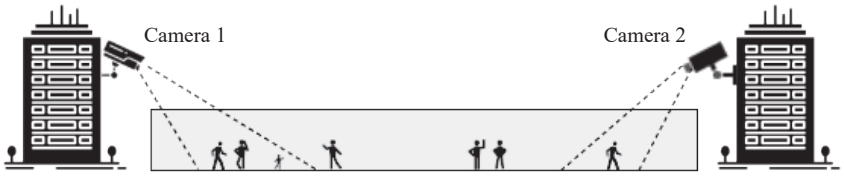

Figure 2. An example of short-term person Re-ID.

When the person of interest walks from camera view one to camera view two, short-term person Re-ID can bridge the gap between these two surveillance views. Given a video sequence (Probe) captured by camera one, person Re-ID tries to identify the same person while crossing the view of camera two. It then publishes a ranked list of images with descending probability of being the person of interest, similar to the primitive person Re-ID framework in Figure 1. Accomplishing this task requires four independent steps: human detection, human tracking, feature extraction, and classification. Together these four steps create a complete person Re-ID pipeline illustrated in Figure 3.

Figure 3. Pipeline of person Re-ID system.

The first two steps of this timeline are independent fields of research which are achieved by numerous methods; however, they are utilised in research to validate the results [8–16]. Feature extraction refers to learning specific properties that make training samples unique, and Classification is the process of matching the feature variables between the training data and the Probe.

Most methods focus on the short-term person Re-ID and use handcrafted or metric features and a single deep neural network solution. In this case, since the area between the two camera views is small, the appearance features will most likely stay the same, and the task of person Re-ID can be achieved by only relying on the appearance or metric features.

Figure 4 shows a typical person Re-ID system with an incorporated human detection and feature extraction step. This system contains a training phase in which a gallery set of feature vectors is generated. This gallery set includes features of all the individuals walking the path. In the testing phase, one descriptor for the person of interest is produced using the same method in the training phase. When the probe person enters the camera network, his feature vector is generated the same way as the training phase, and then compared against the gallery using a classification algorithm or similarity-matching technique. The system will then output the ID of the best-matched re-identified person.

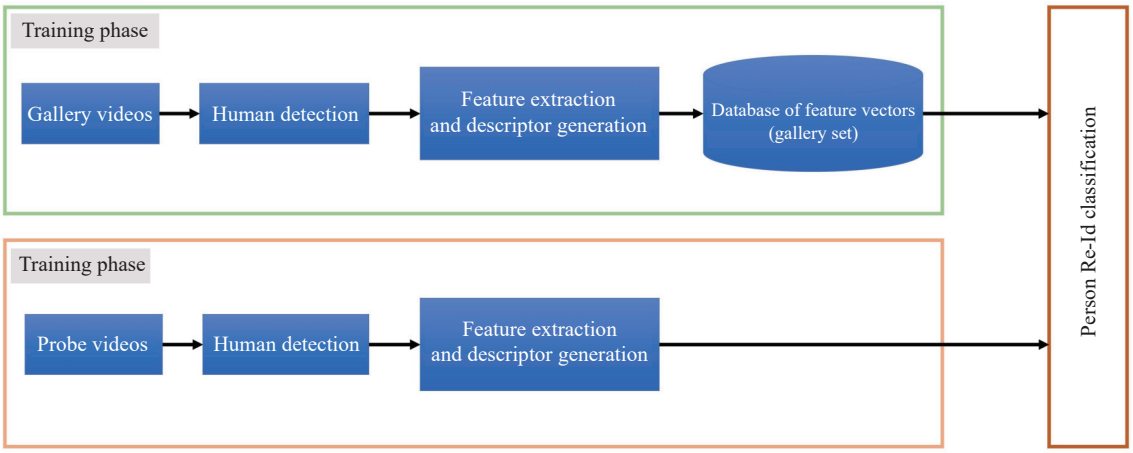

Figure 4. Classic person Re-ID Systems.

Another consideration is the type of sensors used for person Re-ID. Figure 5 shows the most popular sensors used in person Re-ID research. Near-infrared (NIR) cameras are primarily used in situations with low lighting, such as dark indoor conditions or at night [17]. RGB-D sensors or Kinects can acquire depth information which can be helpful for 3D reconstruction or depth perception but is used in very particular person Re-ID applications [18,19]. In practice, most surveillance systems work with the standard RGB cameras, so naturally, researchers focused on reducing the inter-class and intra-class variations on this type of sensors [20,21]. In short term Re-ID, RGB-D cameras are proven to have less performance and more cost than the standard RGB cameras, so they are not practical for widespread use in the real world.

(a) NIR sensor **(b)** RGB-D sensor (kinect) **(c)** RGB Sensor (normal camera)

Figure 5. Most popular sensors for person Re-ID.

Person Re-ID systems could be also characterised based on the type of inputs into image-based and video-based categories. In image-based, the system receives random frames as inputs and focuses on appearance features such as colour and texture since these features remain the same in short periods of time [22–27]. Video-based systems receive a video clip or a collection of successive frames as input. However, in addition to the appearance attributes, video-based systems also explore movement data to improve the performance of the system [28–35].

Short-term person Re-ID methods have achieved great accuracy on publicly available datasets. However, due to several inter-class and intra-class variations, they suffer from low performance and high cost. Subsequently, researchers try to reduce these variations and their effects. In the early years, the main focus was on handcrafting descriptors from images or videos. This is including but not limited to histograms of different colour spaces such as HSV, YCbCr, LAB and LBP which are extracted from overlapping features and then concatenated into single feature vectors, local patterns in binary forms [36–39], a collection of local features [40,41], maximal occurrence representation of local features [25], HOG3D [34], STFV3D [33], features learnt by convolutional neural networks [22,42,43], or a combination of the features above. Metric learning classification methods such as KISSME, local fisher and marginal fisher analysis, top push distance learning, nearest neighbour, NFST and quadratic discriminant analysis are also used to discriminate between the mentioned features [24,25,34,38,44–47].

According to a projection published by IHS markit, there are over one billion surveillance cameras installed around the world as of 2021 [48]. After China, the UK has one of the most substantial numbers of CCTV cameras globally. In 2015 the british security industry association (BSIA) estimated that between four to six million security cameras are installed in the country. London has the highest number of CCTV cameras in the UK. By that estimation, an average Londoner could get caught on camera three hundred times a day [49]. These cameras are used in places from home and shop surveillance to public areas such as airports, shopping centres, metro stations, and other forms of public transportation.

The main reasons behind this radical increase in the use of CCTV are the reduction in the price of cameras and the effectiveness of crime prevention. Utilising CCTV cameras in real-time, the police and security agencies can prevent incidents by detecting suspicious behaviour or gathering evidence such as identifying suspects, witnesses, and vehicles after a crime has been committed. Accordingly, the task of threat detection and person re-identification is left entirely to human operators. These security operators need to possess a high level of visual attention and vigilance to react to rarely occurring incidents. Moreover, many human resources are required to analyse the millions of hours of collected videos, thereby making this task very costly. It is even more taxing and time-consuming to search for a person of interest in thousands of hours of prerecorded videos which requires expert forensics specialists.

Automatic video analysis considerably reduces these costs, and for this reason, it became a critical field of research. This research field tackles problems including but not limited to object detection and recognition, object tracking, human detection, person re-identification, behaviour analysis and violence detection. Solutions to these problems have applications in numerous domains like robotics, entertainment, and to no small extent, video surveillance and security.

Person Re-ID is different from the classic detection and identification tasks since person detection is to distinguish a human from the background, and identification is the process of determining a person's identity in a picture or video. Detection indicates whether there is a person in the provided image, and Identification tells us who it is. However, person Re-ID determines whether an image in a video clip belongs to the same individual who previously appeared in front of the camera.

Usually, the assumption is that the subject wears the same clothing in different camera views, and the appearance stays the same for the re-identification task. This premise produces a significant limitation on the job since people can change their appearance and especially their clothing over the course of days, hours or even minutes. These alterations make re-identification based on appearance unlikely after a certain period of time. The hypothesis is that

biometric features like faces or Iris are not always available in CCTV videos, especially after the rise of the COVID-19 pandemic.

Illumination changes, position variations, viewpoints and inter-object occlusions make appearance-based person Re-ID a notable problem. Most recent models use different features like the colour of clothes and texture to improve their performance. Typically they generate feature histograms, concatenate and finally weigh them according to their importance and distinguishing power [50,51]. These features can be learnt through multiple methods such as boosting, measuring distance metrics and rank learning [52,53]. The downside of these methods is their lack of scalability since the learning process needs constant supervision as the subjects change. It is a better practice not to bias all the weights to global features and give selective weights to more individually unique features such as salient appearance features or gate features such as walking speed and direction and the flow of movement. Human visual attention is studied in [20] and the results imply that attentional competition between features could take place not only based on the global features but also the salient features in individual objects.

To overcome the issues mentioned above, soft biometrics, such as Gait, has been used in the past. Gait is a biometric feature that focuses on a person's walking characteristics and motion features. Other biometrics, such as Iris and face could be altered using a pair of contact lenses or a simple surgical mask after the COVID-19 pandemic. Gait's advantages in video surveillance include the ability to extract features non-invasively from a distance, property acquisition from low resolution, and the ability to extract features even in the dark using different modality cameras. The two key metrics in Gait analysis are spatial and temporal parameters. Spatial and temporal features can describe the state of an object over time or a position in space. Moreover, in the past few years, several publicly available datasets have been published, which can be used to validate our research. In short, working on Gait as a biometric feature for person recognition and re-identification could be a massive help in the war against crime and even prevent terrorist attacks by recognising well-known terrorists and dangerous repeat criminals.

This paper introduces the methods used in Gait person Re-ID and reviews the literature surrounding the subject by dividing the person Re-ID algorithms into three different paradigms.

2. Applications and Challenges

2.1. Applications

Person Re-ID methods have much potential in a wide range of practical applications, from security to health care and even retail. For instance, cross-camera person tracking could understand a scene through computer vision, track people across multiple camera views, perform analysis on the crowd movement and do activity recognition. When the person of interest moves from one camera view to the other, the track will be broken. Therefore, person Re-ID is used to establish connections between the tracks to accomplish cross-camera tracking.

Another form of person Re-ID is tracking by detection, which uses person detection methods to perform tracking of a subject. This task includes modelling diverse data in videos, detecting people in the frames, predicting their motion patterns and performing data association between multiple frames. When person Re-ID is associated with the recognition task, a specific query image will be introduced and searched in an extensive database to perform person retrieval. This method usually produces a list of similar images or frame sequences in the database.

Person Re-ID is also used to observe long-term human behaviour and activity analysis, for example, maximising sales by observing customers' shopping trends, analysing their activities and altering products and even shopping floor layouts. In health care, it can be used to analyse patient behaviour and habits to assist hospital staff with a higher standard of care.

Other applications of the Gait person Re-ID have become more evident in recent years. For example, a considerable portion of human faces were hidden behind masks during the COVID-19 pandemic. Moreover, social or assistive robots carrying out daily collaborative tasks must know whom to reach. Person Re-ID using a biometric pattern such as Gait could replace identification cards or codes for critical infrastructure access control. Brittany which is a biometric recognition tool based on gait analysis and convolutional neural networks (CNNs) is presented in [54]. This could be expanded to autonomous vehicles where 3D motion data could be collected using LIDARs and cameras. Another use of Gait person Re-ID has been presented in [55], which proposes a smartphone-based gait recognition system that can recognize the subject in a real-world environment free of constraints.

2.2. Challenges

2.2.1. Appearance Inconsistency and Clothing Variations

Several challenges must be overcome to solve Person Re-ID's problem and use it in the above applications. Matching a person across different scenes requires dealing with class variations and confusion. The same person can

undergo significant changes in appearance from one scene to another, or two different individuals can have similar appearance features across multiple camera views. These variations include Illumination variation, camera viewpoint variation, pose variation, low resolution, similar clothing, partial occlusion, real-time constraint, clothing change, accessories change, camera settings, a small training set and data labelling cost. Moreover, most models support short-term Re-ID in which they leverage colour and texture and the object carried by the subject. One of the most common challenges of appearance-based methods is the assumption that colours could be assigned to the same object under various lighting conditions, whereas achieving colour consistency under such conditions is not an easy task [56].

Another limitation is appearance consistency, which can only be assumed in a short time span from one camera to another. Since the holistic appearance will change drastically over days and weeks, the technique, as mentioned above, will be ineffective. Practical applications usually require a long-term solution in which the subject's appearance and clothing might have changed after the elapse of a significant period [57]. It is irrefutable that appearance-based short term person Re-ID techniques have made great strides in terms of accuracy in the past years. Still, problems such as massive intraclass variations and inter-class confusion caused by the conditions mentioned above make this a challenging and worthwhile field of research. Some of these challenges can be seen in well-known datasets such as PRID2011 [58], MARS [32], iLIDS-VID [59], DukeMTMC-reID [60].

2.2.2. Insufficient Datasets

Insufficient datasets are another challenge. Several publicly available datasets are out there, but none is large enough regarding the number of camera angles, number of subjects, or recorded period of time. Most datasets are usually recorded using two cameras and a small sample of subjects, and since deep learning models need an extensive training set and validation set, building realistic datasets would help further the progress of person Re-ID research.

2.3. Tackling the Challenges Using Generated Data

When training a model on one dataset and testing on another, more often than not, the performance drops significantly. To overcome this challenge, data augmentation techniques such as generative adversarial network (GAN) [61] have been designed in recent years to introduce large-scale datasets [62,63] or to expand sample data [60,64]. Another important research direction is the unsupervised person Re-ID models. Unlike supervised learning, these models do not train using labelled data in the same environment and have a lower annotation cost [65–67].

Since these features are not robust enough, some authors try to include other modalities, such as depth and thermal data, in their models using deep learning. These models are known as multi-modal methods and are particularly challenging in real life because a framework has to be developed that can handle multiple variations such as subject position and view obstructions [68,69]. Building architectures of deep learning models for person Re-ID is a very time-consuming task, so some researchers are using neural architecture search methods (NAS) to automate the process of architecture engineering [70,71]. The most crucial challenge in NAS methods is that there is no guarantee of how appropriate the CNN would be after NAS has chosen it.

2.4. Tackling Challenges Using Biometric Features

Facing all these problems, long-term approaches using Biometric features have been suggested by researchers in the past. Biometrics is the science of person identification based on physical and behavioural traits like body measurements and calculations related to human characteristics, which can be used to describe and label them [72]. Biometric surveillance identifies a person of interest by extracting features of all the people in the camera network and comparing them to the gallery set. The most commonly used biometrics are categorised as Hard Biometrics and Soft Biometrics. Some Hard Biometrics are fingerprint, iris, face, voice and palm print which do not change over time and are primarily used in access control systems. The acquisition of these biometrics demands a controlled environment and invasive measures. In video surveillance scenarios, people move more freely and without supervision, making it impractical to gather hard biometrics. To tackle the problem of long-term Re-ID, Soft Biometrics [73] like anthropometric measurements, body size, height or Gait are used with better success. These biometrics are more reliable for long-term Re-ID solutions, but also more challenging. Compared to hard biometrics, soft biometrics lack strong indicators of an individual's identity. Nevertheless, the non-invasive nature of Soft Biometrics and the ability to acquire information from a distance without the need for subject cooperation makes them a strong candidate for tackling person Re-ID.

2.4.1. Gait as a Soft Biometric Feature

Among Soft Biometrics, Gait is the most widespread feature for person Re-ID in surveillance networks. Other soft biometrics such as 3D face recognition are also considered, but Gait is more popular because first, it does not require contact with the subject, and the cues gathered from Gait are unique to each individual that are extremely hard

to fake [74]. Gait can also work in low-resolution videos. Gait Re-ID gathers and labels distinctive measurable features of human movement just like any other biometric system. In psychology and neuroscience, focusing on Gait is an essential subject in humans' perception of others. For example, it is a known fact that in cases of prosopagnosia or face blindness, "the patients tend to develop individual coping mechanisms to allow them to identify the people around them, mostly via non-facial cues such as voice, clothing, gait and hairstyle recognition" [75–77]. Moreover, Gait analysis is a valuable tool for the diagnosis of several neurological disorders such as stroke, cerebral palsy, Parkinson's, and sclerosis [78–81].

2.5. Person Re-ID Using Gait

Person Re-ID based on Gait has received substantial attention in the past ten years, especially from the biometric and computer vision community, due to the advantages mentioned above. Gait as a soft biometric can vary with illness, ageing, changes in the emotional state, walking surfaces, shoe type, clothes type, objects carried by the subject, and even clutter in the scene. Moreover, Gait based person Re-ID problem should not be mistaken with Gait based person recognition since they are applied in entirely different scenarios. Recognition is employed in heavily controlled environments, often with a single camera, and the operator can influence conditions including but not limited to background, subject pose, angle of the camera and occlusion. On the other hand, the conditions in person Re-ID are entirely out of control. Since a large camera network is used to solve this type of problem, variables like lighting, number of people and occlusion, and the angle and direction of the walk are unknown. Gait lets us analyse a person of interest from various standpoints and poses. What makes Gait more attractive is the tendency to be used in long-term person Re-ID, relying on more than only spatial signals and appearance features. Gait is considered a type of temporal cue that could provide biometric motion information. Combinations of appearance and soft Biometrics have been used in the past to solve the problem of person Re-ID [24]. On the other hand, the shape of the human body could be considered a spatial signal which can produce valuable information. Several works have tried to extract discriminative features from both spatial and temporal domains [33,82,83].

2.5.1. Model Based vs Model-Free Gait Recognition

Early research on person gait recognition and Re-ID attempted to model the human body because Gait is essentially analysing the motion of distinct parts of the human body [84–86]. Specific characteristics pertaining to various body parts are extracted from each image in the sequence to form a human body model. Then, the parameters will be updated for each silhouette in the sequence and used to create a gait signature. These characteristics are usually metric parameters such as the length and width of a body part like arms, legs, torso, shoulders, heads along with their positions in the images, and all of them can be used to define the walking trajectory or the Euclidean distance between the limbs. Twenty-two such features are introduced in [87,88] where a posture-based method for gait recognition which learned posture characteristics by considering the displacement of all joints between two consecutive frames and a fixed centre of the body coordinate system for all the joints is proposed. A two-point gait representation was introduced in [86], which modelled the motion of limbs regardless of the body shape.

Other works, such as [89], create a motion model by considering the motion of hips and knees in various gait cycle phases. In this approach, the features are extracted from 2D images. Gait parameters, namely the magnitude and phase of the Fourier components, are extracted using a Viewpoint rectification stage. However, because of the low quality of surveillance images captured in the real world, it is implausible to calculate a model robust enough for widespread practical usage.

In conclusion, although large models achieve the best accuracy, they may consume a lot of memory and time when applied to existing video surveillance systems, therefore directly impacting the efficiency. To solve this problem, the trade-off between accuracy and efficiency must be considered in person Re-ID models.

3. Taxonomy of Gait Re-ID Methods

The Re-Identification of humans in digital surveillance systems is discussed in this section, and the current methods employed to tackle this problem are listed. Critically assessing these methods will help exploit the available labelled data from the publicly available datasets more efficiently and improve the training stage's efficiency in irregular Gait person Re-ID. It also shows a significant research gap in areas relating to irregular gait recognition and Re-Identification, which we face in practical scenarios rather than closed lab environments.

It can be assumed that Aristotle in De Motu Animalium (On the Gait of animals) was the modern gait analysis pioneer. A series of papers were published on the biomechanics of Gait for humans under unloaded and loaded conditions by Christian Wilhelm in the 1980s [90].

It is the cinematography's and photography's progression that make the capture of frame sequences (that reveal details of animal and human movements unnoticed by the naked eye) became possible with pioneering work in [91]

and [92]. For example, the horse gallop sequence was normally distorted in painting before the discovery was exposed by aerial photography.

The production of video camera systems in the 1970s helped begin the extensive research and practical application of gait analysis on people with pathological diseases such as Parkinson's and cerebral palsy within a realistic time frame and with low cost. Based on the results obtained by gait analysis, orthopaedic surgery made significant advances in the 1980s. Many orthopaedic hospitals worldwide are currently equipped with gait labs to suggest and develop regimens and plans for doctors' treatment and scheduled follow-ups.

Human identification from a distance or non-intrusive human recognition has gained much interest in the computer vision research community. The Gait recognition study area addresses the identification of a person by the way they walk. Gait recognition proposes quite a few exclusive characteristics compared to other biometric methods. One characteristic that is most attractive to researchers is the unobtrusive nature of science. The subjects do not need to cooperate or pay attention to be identified. Also, no physical information is needed to capture human gait characteristics from a distance.

In this research area, complications may arise when multiple cameras are used to monitor an environment. For example, a person's position is known to us when they are within a single camera view, but problems might accrue as soon as the subject moves out of one view and enters another. This is the essence of the human re-ientification problem. How can the system detect the same person that appeared in another camera view earlier? The purpose of this section is to evaluate the gait-based methods of re-identification.

We classify the past algorithms into three paradigms to best grasp the state-of-the-art methods used in Gait person Re-ID. Figure 6 shows the overall classification of these algorithms. Finally, we will discuss these approaches and their strengths and drawbacks to better understand Gait Re-ID's challenges.

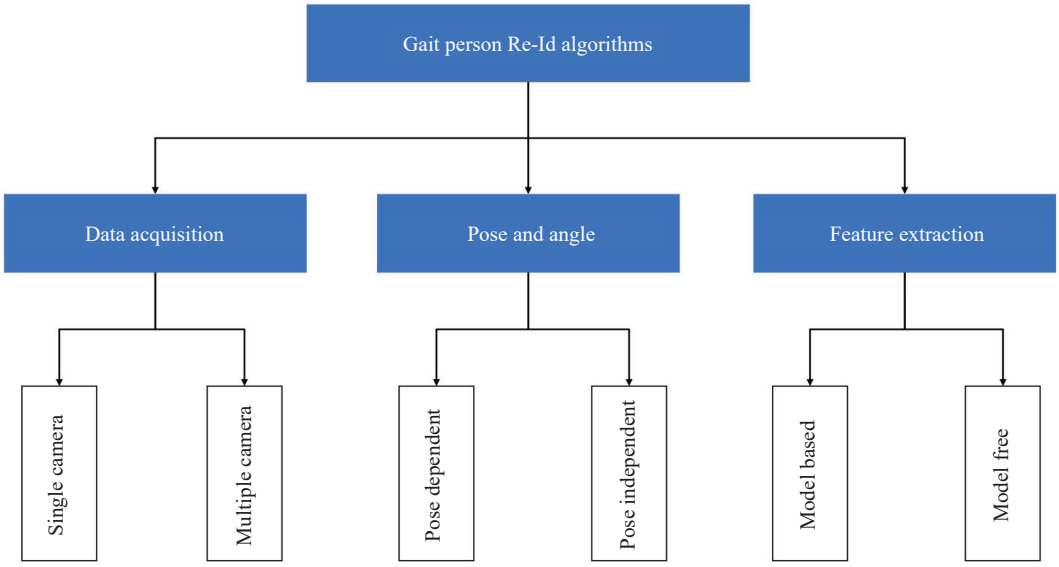

Figure 6. An overall classification of Gait person Re-ID methods.

3.1. Approaches to Data Acquisition

The number and types of cameras used in the raw data acquisition directly impact the person Re-ID algorithm. We can categorise these approaches using a single camera-acquired dataset and those using datasets gathered using multiple cameras.

3.1.1. Single Camera Approaches

Depending on the type of camera used, the algorithm could exploit two-dimensional (2D) or three-dimensional (3D) information. Motion capture (MOCAP) systems or depth sensors such as Kinect can directly acquire a 3D representation of the environment [93–96]. However, datasets generated from 2D cameras are the standard in most previous works [2,89]. A Re-ID scenario usually consists of multiple variations, including camera viewpoints, clothing, lighting and walking speed. When there are multiple cameras in the network, some could be used for the gallery and some for the probe set. Other variations could be used similarly to compare their effect on an algorithm. In most cases, the data used for gait recognition and Re-ID is a short video clip or a sequence of consecutive frames containing gait cycles. Some of the most popular publicly available Gait datasets are listed in Table 1.

Table 1 Details of some of the well-known publicly available datasets for gait recognition

Name	Subjects	Location	Image Dimension	Views	Variations
TUM-GAID	305	Inside	640 × 480	1	Normal Back Pack Shoes
CASIA Dataset A	20	Outside	352 × 240	3	Clothing
CASIA Dataset B	124	Inside	320 × 240	11	Normal Back Pack Coat Lighting
SOTON (Large)	115	Inside/Outside	20 × 20	2	Clothing
CMU MoBo	25	Inside (Treadmill)	640 × 480	6	Slow Walk Incline Walk Ball Walk
OU-MVLP	10,307	Inside	1280 × 980	7	-
OU-ISIR, Treadmill A	34	Inside (Treadmill)	88 × 128	1	Speed
OU-ISIR, Treadmill B	68	Inside (Treadmill)	88 × 128	1	Clothing
OU-ISIR, Treadmill D	185	Inside (Treadmill)	88 × 128	1	Occlusion
i-LIDS (MCT)	119	Inside	576 × 704	5	Clothing Occlusion
iLIDS-VID	300	Outside	576 × 704	2	Lighting Occlusion
PRID2011	245	Outside	64 × 128	2	Lighting Occlusion
KinectREID	71	Inside	Vary	1	-
MARS	1261	Outside	1080 × 1920	6	Lighting
HDA	85	Inside	2560 × 1600	13	-
USF	25	Inside	640 × 480	6	32 Variations
Vislab KS20	20	Inside	Only depth data	1	-

Previous works usually use different combinations of camera views to simulate a real-world scenario even though datasets such as CASIA-B and SAIVT have multiple overlapping and non-overlapping camera views [97]. Only one camera is used for the Probe when running the system, and the rest are used to create the gallery. This approach was used in [98,99] on PKU, SOTON and CASIA, which contain multiple camera viewpoints. The i-LIDS dataset with two different camera views is used in [89]. They used one camera for the Probe and the other for the gallery set. A random fifty-fifty split for each sequence in the dataset regardless of the camera view and tested this approach on iLIDS-VID, HDA+ and PRID2011 datasets is used in [29].

3.1.2. Multiple Camera Approaches

Some previous works used overlapping camera views, but due to constraints in calibration and installation of multiple cameras simultaneously, such works are scarce in the literature. One example of such works is [100], in which they used 16 overlapping cameras to create a 3D gait model of a human. The 3D gait models were generated by the volume intersection of silhouettes extracted from walking frame sequences. Then, random viewpoints from these synthetic images were chosen to create the gallery. Using kinect and MOCAP system will considerably reduce the amount of computation and complexity of the problem since they can acquire the 3D data directly and use the same camera for both Probe and gallery sets [95−101,103]. In these works, people walk in arbitrary directions or side to side in front of the same Kinect 3D sensor. MOCAP systems for person Re-ID is used in [93, 94]. In [93], they used the dataset of Carnegie Mellon University. This dataset was collected using 12 vicon infrared MX-40 cameras with people wearing marked black clothes. The markers were only visible in infrared and were used to produce the 3D information.

Although in the past few years, there was a paradigm shift towards methods with overlapping viewpoints with kinect or MOCAP systems. Nevertheless, there is still an overwhelming amount of research on 2D surveillance networks since the use of such sophisticated devices (which can acquire direct 3D motion information) is not yet incorporated in the real world.

3.2. Approaches to Pose and Angle

In automatic surveillance systems, the pose of a person or human pose is determined by the direction of the walk and the camera viewpoint. Depending on the pose, the dynamic and static information acquired from the image sequence can change when switching between camera views or when a period of time has passed. Conditional on the human pose, pose-dependent and pose-independent approaches are reported in the literature. In pose-dependent, the data applied to the person Re-ID system is restricted to only one direction or camera view, but in pose-independent,

any arbitrary viewpoint or direction could be used as the input of the system.

3.2.1. Pose-Dependent Approaches

Pose-dependant approaches are most useful for indoor scenarios where walking directions will not change in relation to the camera, such as station entrances and shopping centre corridors. Numerous publicly available datasets, including iLIDS-VID, CASIA, PRID2011 and TUM-GAID, support the pose-dependant approach. Single camera viewpoint approach has been used in [102−104]. In [102] both side and top views are used, and in [103,104] they use frontal view, but most past research is focused on the lateral (side) human view because it enables a more clear observation of the human Gait and provides a consistent amount of self-occlusion in each frame [83,105]. Some works choose the system inputs based on the viewpoints provided by the dataset [98,99] for example, lateral for CASIA and TUM-GAID and frontal and back for PKU.

3.2.2. Pose-Independent Approaches

Pose-independent or cross-view are among the most practical methods for real-world applications because the human walk's direction is most likely arbitrary in an uncontrolled data acquisition setup. The viewpoint variation puts a higher computational strain on the person Re-ID system. On the other hand, the input data must be of a higher quality than approaches dependent on only one view. The gallery set in these approaches is constructed from random data collected from arbitrary walking directions, which will be put against a random probe set for testing. For example, [99] produces the gallery using all the 11 viewpoints in CASIA Dataset B and then tests the algorithm by choosing a random probe image from the same datasets. The same technique is used in [29,106] and [2,89], where the gallery set is created from random viewpoints. A view transformation model (VTM), which exploits projection techniques to tackle pose variation problems is provided by [89]. Models such as VTM are utilised to transform multiple data samples into the same angle. Similarly, sparsified representation-based cross-view model and discriminative video ranking model were used by [97] and [29], respectively.

The recent RealGait model [107], which sets a new state-of-the-art for gait recognition in the wild, is a pose-independent deep network model for cross-scene gait recognition. In this paper, they constructed a new gait dataset by extracting silhouettes from a real person Re-ID challenge in which the subjects are walking in an irregular manner. This dataset can be used to correlate gait recognition with person re-identification.

Furthermore, 3D data can reduce the computational cost of view alignment using simple geometrical transformation like [100], which creates 3D models from 2D images that multiple overlapping cameras have acquired. Synthesising these models generated virtual images with random viewpoints and constructed a gallery. A probe was chosen from authentic images collected by a single-view camera to test the algorithm. MOCAP was also employed by [94] and [108] to collect 3D data for a pose-independent person Re-ID system. Since kinect can provide a 3D volumetric representation of the human body, some works such as [95] use skeleton coordinates provided by kinect to demonstrate the impact of viewpoint variation on Gait person Re-ID systems. They use these findings in [101], and [109] in which they proposed a context-aware Gait person Re-ID method that used viewpoints as the context in their study.

3.3. Approaches to Feature Extraction

3.3.1. Human Gait Cycle

Gait signature or Gait feature is the subject's unique characteristics extracted from a sequence of sample images over a Gait cycle or stride. To understand the Gait cycle, we need to analyse, isolate and quantify a unique short and repeated task when someone walks. A Gait cycle can be measured from any gait event to the same gait event on the same foot. However, it is implied in the literature that a Gait cycle should be measured from the strike of one foot to the ground to the next strike of the same foot in a person's walking pattern. The Gait cycle is considered the fundamental unit of Gait. By measuring its temporal and spatial aspects, we can extract Gait signatures for a particular person of interest.

There are two primary phases of Gait cycle: the "swing" phase and the "stance" phase. These phases alternate and repeat when a person walks from point A to point B. The stance phase is the duration when the foot is on the ground, and the swing phase is the whole time when the foot is in the air. Consequently, observing the movements of two lower limbs is imperative in extracting spatial and temporal features. A breakdown of a gait cycle is presented in [110]. As the paper illustrates, when both limbs are in the stance phase simultaneously, the legs are in bipedal or double support, and when only one leg is in the stance phase, it is in the uni-pedal or single support sub-phase. According to [111], the swing phase has four sub-phases. (1) Pre-swing, which is when the foot is pushed off the ground, and the transition between the stance phase and swing phase happens, (2) Initial swing, when the foot clears off the ground. (3) Mid swing in which advancement of the foot continues, and (4) terminal swing that the foot is back in the

beginning position of the gait cycle on the ground.

The traditional feature extraction algorithms in the literature are divided into two categories: model-based and model-free (motion-based) approaches. The model-based methods use the kinematics model of human Gait, and the model-free or motion-based approach finds correspondence between sequences of images (optical flow or silhouette) of the same Person to extract a gait signature. Table 2 shows examples of different works using model-based and model-free approaches. Another feature extraction category is based on deep learning, which has become more prevalent due to better performance and less complexity in discovering gait cycles in real-world situations.

Table 2 Details of some traditional methods in the literature

Method	Approach	Feature	Feature Analysis Technique
[93]	Model-Based	3D joint info	MMC PCA LDA
[89]	Model-Based	Motion model	Haar-like template for localization and magnitude and phase of fourier components for gait signature and KNN classifier
[104]	Model-Free and Model-Based	Fusion of depth information	Soft biometric cues and point cloud voxel-based width image for recognition; LMNN classifier
[33]	Model-Free	2D Silhouettes	Gait and appearance features combined
[105]	Model-Free	Silhouettes	STHOG and colour fusion
[103]	Model-Free	Optical flow	HOFEI
[101]	Model-Based	3D joint info	Context based ensemble fusion and SFS feature selection
[106, 29]	Model-Free	2D Silhouettes	ColHOG3D
[98]	Model-Free	3D joint info	Swiss system based cascade ranking
[83]	Model-Free	2D Silhouettes	Virtual 3D sequential model generation
[97]	Model-Free	2D Silhouettes	GEI-FDEI HSV
[100]	Model-Free	2D Silhouettes	Virtual 3D sequential model generation
[102]	Model-Free	Point cloud	Frequency response of the height dynamics

3.3.2. Model-Based Approaches Based on the Human Body

Early research on Person gait recognition and Re-ID attempted to model the human body because Gait essentially analyses the motion of distinct parts of the human body [84–86]. In these approaches, the silhouettes are obtained from 2D images by background subtraction and Binarisation. Specific characteristics pertaining to various body parts are extracted from each image in the sequence to form a human body model. Then, the parameters will be updated over time for each silhouette in the sequence and used to create a gait signature. These characteristics are usually metric parameters such as the length and width of a body part like arms, legs, torso, shoulders, head and their positions in the images. They can define the walking trajectory or the Euclidean distance between the limbs. Twenty-two such features are introduced in [87]. A posture-based method for gait recognition which learned posture characteristics by considering the displacement of all joints between two consecutive frames and a fixed centre of body coordinate system for all the joints is proposed in [88]. A two-point gait representation was introduced in [86], which modelled the motion of limbs regardless of the body shape. Other works such as [89] create a motion model by considering the motion of hips and knees in various phases of a gait cycle. In this approach, the features are extracted from 2D images. Gait parameters, namely the magnitude and phase of the Fourier components, are extracted using a Viewpoint rectification stage.

Because of the low quality of surveillance images captured in the real world, it is implausible to calculate a model robust enough for widespread practical usage. In recent years depth sensors (i.e. Kinect) and MOCAP systems made modelling the human body more manageable. With the help of a Kinect [112] extracted comprehensive gait information from all parts of the body using a 3D virtual skeleton. Other works such as [104] employed this method in gait recognition and automatic Re-ID systems. Two kinects RGB-D cameras were used in [104] to acquire information from the person's frontal and back view. Since kinects have a limited range for sensing depth, they each captured only a part of a subject's gait cycle, so they fused the information from all the sensors. The authors used the kinect software development kit to compute a set of soft biometric features for person Re-ID when the subject moves from one kinect to the next, and then feature vectors called width images were constructed at the granularity of small fractions of a Gait cycle.

3.3.3. Model Free Approaches

The model-free approaches can again be categorised into sequential and spatial-temporal motion-based meth-

ods. Sequential Gait is presented as a time sequence of the person's poses. The sequential model-free approach was first proposed in [113] in which temporal templates represent the motion. The temporal templates spot the motion and record a history of these movements. In a spatial-temporal model-free approach, Gait is represented by mapping the motions through time and space [114]. Different spatial-temporal methods are proposed in the literature, and almost all of them use silhouette sequences. Gait energy image (GEI), first introduced in [115], is a spatial-temporal representation of Gait that characterises the human walk properties for individual gait recognition, where a single image template is produced by taking the average of all binary silhouettes over the entire gait cycle.

The authors in [115] show how GEIs are generated from sequences of human walk silhouettes. For binary gait silhouette images $B_t(x,y)$, the Gray level GEI is computed using the below formula where N is the number of frames in a sequence, t is the moment in time (frame number), and (x,y) are the position coordinates in a 2D image. GEIs and their variations became the primary feature for many gait person Re-ID research, including [97]. In this work, GEI and its variation called frame difference energy image (FDEI) were used as gait features. Other works such as [103], and [83] also used histogram of flow energy images (HOFEI) and pose energy images, respectively. In [33] gait energy images and appearance features such as HSV histograms were fused to deliver a spatial-temporal model [115].

$$G(x,y) = \frac{1}{N}\sum_{t=1}^{N} B_t(x,y), \tag{1}$$

Optical flow is also used as a feature in several past works. Optical flow images (GFI) were proposed in [116] for the first time in 2011. The basis of GFIs was also sequences of binary silhouette images. The process of GFI generation was depicted and presented in [116]. Variations of optical flow-based methods appeared in the literature in the next years. A pyramidal fisher motion for multi-view gait recognition was used in [117] as descriptor for motions based on short-term trajectories of points. Other methods such as [118] and [119] are also based on silhouettes. A partial similarity matching method for Gait that could construct partial GEIs from 2D silhouettes was proposed in [118]. In addition, [119] generated virtual views by combining multi-view matrix representation and randomised kernel extreme learning. Spatial-temporal histogram of oriented gradient (STHOG) is proposed in [105] where it was suggested that the STHOG feature can represent both Motion and shape data. A syntactic 3D volume of the subject from images acquired by multiple overlapping 2D cameras from which features were extracted for gait recognition was generated in [100].

3.3.4. Fusion of Modality Approaches

Re-ID results are improved by combining multiple features. Gait-based features have been successfully combined with other types of features in Re-ID.

3.3.5. Deep Learning-Based Approaches

Deep learning is a subarea of machine learning (ML) that tries to train computers on learning by example. This technique has been used before and is the key to technologies such as driverless cars, detecting a stop sign or distinguishing an object from a human being. Other applications can be in voice control devices such as google assistant, Apple Siri and Amazon Alexa projects. Deep learning is getting much attention from the research community because its results were not previously possible.

In this method, a computer model learns how to perform classifications using the training it previously got from sound, text, images or video frames and gains unbelievable accuracy that is sometimes better than humans. Vast collections of labelled data and neural networks with various levels are used to train deep learning models. Deep learning receives such impressive recognition because of its higher than before accuracy level. This helps big companies keep the users happy, meet their expectations, and decrease safety concerns in more critical projects like driverless cars. Recent developments show that deep learning transcends human beings in tasks like feature extraction in images.

In recent years deep learning approaches to gait recognition have been progressing fast [120,121]. The idea of a siamese convolutional neural network (SCNN) was introduced in [122] and utilised in numerous research after that [123−125]. A Siamese network uses multiple identical sub-networks with the same weights and parameters to find a relationship between two related things. By mapping Input patterns into a target space and calculating their similarity metrics, they can discriminate between objects. If the objects are the same, the metrics will be small, and if they are different, the metrics are significant. Some applications of Siamese networks are face recognition, signature verification, and in recent years person re-identification [125−127]. Siamese frameworks are suitable for gait recognition scenarios since there are usually many categories and a small volume of sample data in each category. Siamese CNN for gait feature extraction and person Re-ID was first mentioned in [120] and then in [128]. Based on that [121] pro-

posed a CNN-based similarity learning method for gait recognition, and [129] designed a cycle-consistent cross-view gait recognition method by using generative adversarial networks (GANs) to create realistic GEIs. To handle angle variations, authors in [119] proposed localised Grassmann mean representatives with partial least squares regression (LoGPLS) method for gait recognition, and authors in [130] offered an autocorrelation feature which is very close to the GEI. Multi-channel templates for Gait called period energy images (PEI) and local gate energy images (LGEI) with a self adaptive hidden markov model (SAHMM) were introduced in [131] and [132], respectively.

All of the above works used GEIs as input data for their systems. As mentioned before, in the process of generating GEIs from gait sequences, a vast amount of Gait temporal data will be lost, so in more recent works, attention mechanisms [133,28], and pooling approaches [134,135] were used. A sequential vector of locally aggregated descriptor (SeqVLAD) was proposed in [133] which combined convolutional RNNs with a VLAD encoding process to combine spatial and temporal information from videos. Sparse temporal pooling, line pooling and trajectory pooling were also used in various works to extract gait features [134,135]. Some works, for example, [136] and [137] used raw silhouette sequences instead of GEI or its variations to preserve the temporal features. In [137] they used ResNet and Long-Short Term Memory (LSTM). GANbased methods were used in recent works on large-scale datasets [130,138,139]. Researchers in [140] even used RGB image sequences instead of silhouettes with autoencoder and LSTM networks. Liao et al. [141] even proposed a model-based gait recognition method with the use of CNNs, which did gait recognition as well as predicting the angle. A list is provided in Table 3 of some more important deep learning methods along with their features and techniques.

Table 3 Details of most notable deep learning methods in the literature

Method	Feature Analysis Technique	Feature
[141]	Convolutional Neural Networks (CNNs)	Joints Relationship Pyramid Mapping (JRPM)
[131]	Generative Adversarial Networks (GAN)	Period Energy Image (PEI)
[137,140]	Convolutional Neural Networks (CNNs) Long Short-Term Memory (LSTM)	Silhouette Sequences
[132]	Self-Adaptive Hidden Markov Model (SAHMM)	Local Gait Energy Image (LGEI)
[136]	Convolutional Neural Network (CNN)	Silhouette Sequences
[138]	Generative Adversarial Networks (GAN)	Gait Energy Image (GEI)
[119]	Localized Grassmann mean representatives	Partial Least Squares Regression (LoGPLS)
[121,130]	Localized Grassmann mean representatives	Convolutional Neural Network (CNN)

It is essential to mention that various features can be fused to improve a person Re-ID system's performance. However, noise and irrelevant information have to be eliminated. Many feature selection and dimension reduction techniques have been carried out in the past to eliminate redundant and irrelevant data collected in feature extraction [142]. Principal component analysis (PCA) is one of the most popular techniques for dimension reduction [83, 99]. Moreover, probabilistic principal component analysis (PPCA) [143] and multilinear principal component analysis (MPCA) [94] have been used in person Re-ID. The authors in [93] even used a fusion of PCA and linear discriminant analysis(LDA) to achieve better accuracy. Recherches in [101] and [102] utilised algorithms such as KL-divergence and Sequential Forward Selection to achieve feature selection.

Although deep learning was theorised in the 1980s, it was not recognised until recently due to two main reasons. Firstly, deep learning algorithms need large quantities of labelled footage or other data. For example, many subjects must be used for training in human recognition before testing the model's accuracy. Additionally, deep learning algorithms need significant computing power. The advancements in parallel architectures for high-performance graphical processing units (HPGPUs) help this issue immensely. These architectures can be combined with clusters or cloud computing to reduce the training time significantly.

Despite the disadvantages mentioned above, there are practical implementations of deep learning in industries such as (1) Automated driving where objects like pedestrians and stop signs are automatically detected, which reduces the possibility of collisions. (2) Medical research, specifically cancer research where deep learning is used to detect cancerous cells in a human body automatically. UCLA researchers build a microprocessor that can detect and analyse 36 million images per second using deep learning and photonic time stretch for cancer diagnosis. (3) Aerospace and defence projects where deep learning satellites are used to identify objects in areas of interest and safe zones for troops deployed in a specific area. (4) Industrial automation, where the workers' safety is improved by detecting the unsafe distance (from the machines) for workers operating heavy machines. (5) Electronics for speech and voice recognition such as voice-assisted tools that translate speech into words or control devices around the house.

3.4. Multiple Modality Spatial-Temporal Approaches

Fusion of modalities have been used in the past for handcrafted features, optical flow maps, grayscale and silhouette image sequences to improve the robustness of gait recognition methods [144]. Similar feature fusion could be performed on different combinations of modalities and networks like in [145]. In the past, [146] has proved the most promising results for this task. Several types of input modalities are also considered to generate a single gait signature from the input sequences, including handcrafted high-level descriptors, grayscale images, silhouette sequences, and optical flow maps [147]. Moreover, information fusion at different levels has been studied for the final two-stream spatial-temporal CNN architecture. The most challenging part of this approach was consolidating temporal features and preserving them as well as the spatial features for use in high-layered networks without losing information along the way. All modalities were tried out on the range of our proposed architectures and compared against state-of-the-art [148], where attempts were made to fuse two parallel networks at different architectural levels with various fusion techniques and appropriate modalities to improve the results. A very high-level example of this proposed approach can be seen in Figure 7. As can be seen, image sequences of a particular person are fed into the architecture to extract features from two separate modalities, namely optical flow (OF) and grayscale. Using different CNN architectures, they extract a gait signature based on each modality and fused them at the end for use in a classifier.

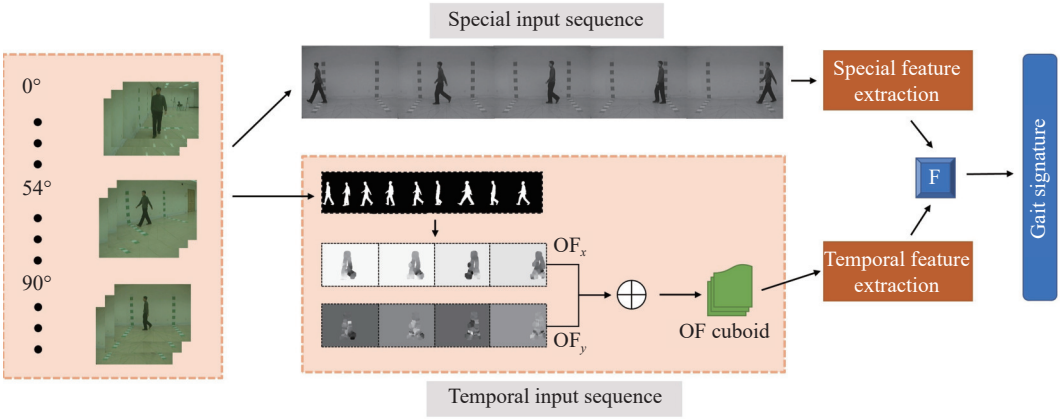

Figure 7. High-level illustration of our proposed two-stream CNN.

3.4.1. Use of High-Level Descriptors

It is not possible to use low-level descriptors as inputs for an algorithm since they only provide an abundance of information about corners and edges, and in the case of divergence, low-level descriptors only provide information of curl, sheer (DCS) [149] and low-level kinematic motions. Such information is usually summarised as inputs for a support vector machine (SVM) or other classification algorithms.

To find the patterns in the raw data, a clustering method could be employed to create a dictionary of gait signature vectors using the collected patterns. When creating the high-level descriptor, no visual features are considered since the algorithm uses low-level features as input data. Authors such as [150] used fisher vectors (FVs) to create hybrid classification architectures with the help of CNNs. A pyramidal fisher Motion descriptor for Multi-view Gait Recognition was proposed in [117] which had promising results on the SVM classification algorithm. The descriptor uses densely collected local motion features as low-level descriptors and fisher vectors to summarise low-level features so as to create high-level descriptors for the SVM, where an extension of Bag of Words (BOW) [151] was used based on the gaussian mixture model (GMM), and an image representation was computed by a gradient vector using a generative probabilistic model in fisher vectors encoding. In [117], the term fisher motion was used to refer to the high-level descriptor generated from low-level motion descriptors.

To create high-level descriptors according to the above method, a video clip (V) can be represented by the below gradient vector equation for T low-level descriptors (x_t), where, $p(x|\lambda)$ is the low-level descriptor independently generated by the Gaussian Mixture Model with $\lambda = \{w_i, \mu_i, i = 1...N\}$ parameters and ∇_λ is the gradient operator.

$$G_\lambda(V) = \frac{1}{T}\sum_{t=1}^{T}\nabla_\lambda log p(x_t|\lambda) \tag{2}$$

With this method, a fisher kernel is calculated to compare two video clips V and W. F_λ is the Fisher information matrix described in [152]. As F_λ is symmetric and positive, it has a Cholesky decomposition $F_\lambda = L_\lambda^T L_\lambda$ and $K(V, W)$ can be rewritten as a dot-product between normalized vectors which is then known as the Fisher Vector of

video V.

$$K(V,W) = G_\lambda(V)^T F_\lambda G_\lambda(W) \tag{3}$$

In this method, the training set is used to compute a dictionary of patterns obtained with the Gaussian Mixture Model, and then a gradient vector is computed in the dictionary to build a feature vector used in a classification algorithm.

3.4.2. Use of Optical Flow

Features can be extracted based on the assumption that temporal descriptors for Gait could be extracted from optical flow [153] and the fact that CNNs can self-learn features from optical flow maps. Optical flow is essentially the movement of a person from one frame to the next in a sequence of frames. The motion is calculated based on the movement between the camera and the Person, giving us valuable temporal information in a gait sequence. A person's movements can be tracked across consecutive frames and estimate their position and even walking velocity using optical flow. Image intensity I is the basis for calculating optical flow. The optical flow process between two proceeding frame, where image intensity was represented as a function of time and space was presented in [116]. Here, t is time or frame number and (x,y) is the position of one pixel in the 2D image. If the same pixel id is displaced by d_x in the x direction and d_y in the y direction in a period of t, then the new image could be described as:

$$I(x+d_x, y+d_y, t+d_t) \tag{4}$$

Hence, the optical flow equation is shown in Euation 5 Where $\frac{\partial I}{\partial x}$ is horizontal image gradient along the x axis and $\frac{\partial I}{\partial y}$ is the vertical image gradient along the y axis and $\frac{\partial I}{\partial t}$ is the temporal image gradient. To solve the optical flow equation and determine the motion over time, one just needs to solve $u = \frac{d_x}{d_t}$ and $v = \frac{d_y}{d_t}$. As multiple unknown variables are needed to solve this equation, some algorithms have been proposed to address the issue.

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \tag{5}$$

Optical flow has been used in the past as the input for CNN and archived excellent results for action recognition [154−156]. There are two types of optical flow, namely Sparse-OF and Dense-OF algorithms. The output of a typical Sparse-OF algorithm is vectors which contain information about edges and corners and some other features of the moving object in the frame. In contrast, Dense-OF algorithms produce vectors for all the pixels in the image like the Farneback algorithm [157]. Some traditional implementations of sparse optical flow are Horn-Schunck [158] and Lucas-Kanade [159] algorithms. Most of the techniques rely on energy minimization in a coarse-to-fine framework [160,161] and [162]. A numerical method warps one of the images towards the other from the most coarse level to the finest level and cleans the optical flow with each iteration, however, for motion estimation, normal flow-based methods present a better outcome [163,164]. Deep learning methods such as FLowNet [165], FlowNet2 [166] and LiteFlowNet [167], use CNNs to estimate the optical flow and are the most promising in action recognition problems. An approach was proposed in [148], where they use a stack of optical flow displacement fields for several consecutive frames, much like proposed in [154].

3.4.3. Fusion

Depending on the number of modalities in the above approaches the extracted features by the CNNs have to be combined using fusion to extract a single gait signature. The expectation is that a better classification score can be achieved by fusing more modalities. Two popular fusion methods could be used for this purpose: (1) late fusion and (2) early fusion.

• Late Fusion

It happens after the softmax layer to combine the output of all the CNNs. For this purpose, we can take the product or the sum of all softmax vectors extracted at the end of each CNN. So if we have n probability vectors at the end of each CNN for m different modalities (two modalities at this point), we can use any of the two equations below to produce a result vector that shows the probability of the Person in the sequence s having the same identity as the Person in class c. The symbol $0 < \rho < 1$ represents a weight related to the modality and is assigned based on experiments.

$$Result = \prod_{i=1}^{n} P_i(m_i = c) \tag{6}$$

$$Result = \sum_{i=1}^{n} \rho P_i(m_i = c) \tag{7}$$

The two-stream architecture in [154] employed late fusion, but since the fusion happens using the softmax layers, it neglects the correlation between temporal and spatial features at a pixel level. Also, since spatial CNN works on one image at a time and temporal CNN on a stack of ten optical flow images, a lot of the temporal information is ignored.

• Early Fusion

On the other hand, early fusion can happen at any layer of the CNN architecture as long as it is before the softmax layer. If the fusion is performed in a convolutional layer, the descriptor is a matrix, and if it is performed on a fully connected layer, it is a vector. Fusing at a convolutional layer is only possible if the two networks have the same spatial resolution at the location of the layers intended to be fused. We can stack (overlay) the layers from one network to the other so the channel responses at the same pixels in the temporal and spatial stream can correspond. If we presume that (1) separate channels in the spatial stream are responsible for different parts of the human body (leg, foot, head), and (2) a channel in the temporal stream is in charge of the motion information (moving the foot forward), then after stacking the channels, the kernels in the next layer needs to learn this correspondence as weights.

There are several ways of fusing layers between two networks. If f is a fusion function that fuses X_t^a and X_t^b which are two feature maps at time t, then f has the output of $y_t \in \Re^{H'' \times W'' \times C''}$ for $x_t^a \in \Re^{H \times W \times C}$ and $x_t^b \in \Re^{H' \times W' \times C'}$. The width, height and channel numbers are represented as W, H and C respectively and the fusion function is represented as:

$$f : X_t^a, X_t^b \Rightarrow y_t \tag{8}$$

To decide where and how to perform the fusion, we first go through a range of feasible methods introduced in the literature. We assume the same dimensions for X_t^a, X_t^b and y_t as we discussed before. So for $x^a, X^b, y \in \Re^{H \times W \times D}$, sum fusion sums the feature maps at the same location in space indicated by i, j and, where c is the number of channels.

$$y_{(i,j,c)} = X_{(i,j,c)}^a + X_{(i,j,c)}^b \tag{9}$$

Since the channel numbers are chosen at random, the correspondence between the networks is arbitrary. Therefore, more training is necessary over the following layers to make this correspondence helpful. The same rule applies to MAX fusion which takes the maximum of two feature maps.

$$y_{(i,j,c)} = max\{X_{(i,j,c)}^a, X_{(i,j,c)}^b\} \tag{10}$$

Concatenation fusion stacks the feature maps at location (i, j) for channels c and outputs $y \in \Re^{H \times W \times 2C}$, but it does not define correspondence, so they need to be defined in the next layers by learning new kernels.

$$y_{(i,j,2c)} = X_{(i,j,c)}^a \quad \text{and} \quad y_{(i,j,2c-1)} = X_{(i,j,c)}^b \tag{11}$$

The feature maps can be stacked as shown and succeeded by a convolution on the result with a kernel $k \in \Re^{1 \times 1 \times 2C \times C}$ and bias of $b \in \Re^C$, to solve the problem of concatenation fusion. Note that the number of output channels is C and the dimensions for the kernel are $1 \times 1 \times 2C$. For early fusion in the fully connected layers of 2D and 3D CNNs and also fusion of the gait signatures obtained by each modality, we perform a concatenation operation followed by another three fully connected layers with dropout before the softmax layer. This fusion is best performed after the last fully connected layer since the signature is already extracted at that point. In ResNet, we will not add fully connected layers after concatenation since it might lead to overfitting. Note that the Relu function is used for all activations in our networks, so for the extra layers, it remains the same.

$$Conv f = y * k + b \tag{12}$$

Adding fusion significantly affects the number of parameters used in the computation. Moreover, we also need to consider the fusion process over time (temporally). Our first consideration is to average all the feature maps x_t over time t which was used in [154]. By this way, only two-dimensional pooling is possible, which loses lots of temporal information. Next, we consider using 3D pooling [168] on a stack of feature maps $x \in \Re^{H \times W \times T \times C}$ over time t. This layer applies max pooling to a stack of data of size $\Re^{W' \times H' \times T'}$. It is also possible to perform a convolution with a fusion kernel before performing 3D pooling, much like [169]. This fusion technique is illustrated in [148].

3.5. Spatial-Temporal Attention Approaches

As video surveillance becomes more widespread and video data becomes more available, the need for a better and more robust person Re-ID framework becomes more evident. Furthermore, pedestrians in surveillance videos are

the central area of concern in real-world security systems. Hence, new challenges are presented to the research community to identify a person of interest and understand human motion. Our study on the human Gait provides a non-invasive and robust way of feature extraction which can be used for gait recognition in surveillance systems in a remote and non-invasive manner.

In real-world scenarios, the subjects passing through an elaborate surveillance network cannot be expected to act predictably. We might not even get a complete gait cycle from a person of interest in most cases. Clothing variation or carry bags considerably impact the system's performance in re-identification problems. Other abnormalities, including the camera angle, significantly aggravate the intra-class variation. Moreover, the similarity between the gait appearances of different people extracted from low-level information introduces inter-class variations, resulting in similar gait signatures in more complex cases.

Unfortunately, most existing methods use shallow motion cues by employing GEIs to represent temporal data, leading to the loss of a vast amount of dynamic information. Although it has been tried in the literature [120,121], [170] to create a deep learning-based gait recognition option that can robustly extract gait features, at least one complete gait cycle needs to be detected, and this is not robust to the viewpoint changes. Therefore, irregular gait recognition concerning viewpoint variations still need particular attention.

A profound challenge in computer vision is detecting the salient regions of an image [171]. These approaches use spatial-temporal attention mechanisms to learn gait information from a sequence of images. Humans usually focus on distinctive features in a person's movements and distinguishing characteristics to recognise one another. In other words, their attention is diverted to specific regions in a scene to find salient features. Numerous works in the literature focus on directing the network's attention with saliency maps [171–173]. It has also been used in the past for action recognition problems like [174] which uses selective focus on RGB videos. A spatial–temporal summary network for feature learning in Irregular Gait recognition was proposed in [175]. For irregular gait recognition, their model learns spatial-temporal and view-independent features. In order to represent the periodic motion cues of irregular gait sequences, they designed the gate mechanism with attentive spatial-temporal summaries. With the spatial feature maps, the general attention and residual attention components can focus on semantic regions that discriminate identity from other semantic regions. As part of the proposed attentive temporal summary component, adaptive attention is automatically assigned to enhance discriminative gait timesteps and suppress redundant ones.

3.5.1. Attentional Interfacing

When watching a video clip, we focus on some frames more than others. Depending on what we are looking for, there might be areas of each frame that attract more attention. This behaviour is also extended to other applications that involve a continuous sequence such as transcription, translation [176], parsing text [177] or voice recognition [178]. While transcribing an audio recording, we would listen to the section we are writing down more carefully.

By using an attention mechanism, neural networks could mimic the same behaviour by focusing on essential sections of the information sequence. This can be achieved using a recurrent neural network that watches over another RNN using an attention distribution interface. The attending RNN can focus on different positions of the attended RNN at each timestep by regulating its focus to different extents. This way, the RNN could learn where to focus deferentially. The attention distribution in such systems is usually based on content. The attending RNN generates a query. The query is combined (dot product) with each input item to generate a score. These scores are usually fed into a softmax layer to generate the attention distribution.

Another example of interfacing CNN and RNN with an attention mechanism is diagnosing diseases from medical images. In [179], the authors used an attention mechanism to diagnose thorax disease from chest x-rays. Their attention-guided convolutional neural network (AG-CNN) consists of three branches. The global and local branches use ResNet-50 as a backbone. A fusion branch combines the pooling outputs of the global and local branches. In [180], the input image was encoded by a dense CNN and decoded using an LSTM network to capture mutual dependencies between the labels.

3.5.2. Soft Attention vs Hard Attention

There are two types of attention known as "Hard Attention" and "Soft Attention". The main idea behind attention is to change the weight of certain features according to some externally or internally provided weights. When these weights are provided internally, the process will be called "Self Attention". Soft attention uses continuous weights, and hard attention uses binary weights. For example, [179] is an example of hard attention since it crops part of the image using its global branch. So the cropped section weights one, and the rest of the image weights zero. The downside of hard attention is the fact that it is none-differentiable(In calculus, a differentiable function of one real variable is a function whose derivative exists at each point in its domain. In other words, the graph of a differentiable function has a non-vertical tangent line at each interior point in its domain. A differentiable function is locally well

approximated and smooth as a linear function at each interior point and does not contain any break, angle, or cusp.), therefore unable to train end-to-end.

Soft attention is used in several works in order to train attention weights. One example of soft attention is "Squeeze and excitation blocks" introduced in [181], which re-weight the responses at certain levels of the network to model dependencies between the channels of the features extracted by the CNN. After each convolution, the squeeze operation aggregates the global feature responses at a spatial level, and then the excitation operation produces a channel-wise weight response. Finally, the convolution operation and the excitation operation outputs are multiplied (channel-wise) and passed into the next convolutional layer. These blocks include the global information in the network's decision-making by accumulating the information from the entire receptive field. However, a typical convolution only looks at local spatial information. The excitation operation's generated weights are similar for different classes in lower layers of the network, but they become more discriminative in higher layers. An important point to notice is that most excitation weights become one at the last stage of the network, so the squeeze and excitation blocks should not be used at this level. Moreover, these blocks can be integrated into popular networks such as ResNet at particular stages without considerable overhead to the training parameters.

3.5.3. Self-Attention

Self-attention (intra-attention) finds relationships between different positions of a single sequence. This technique is used in machine reading with the help of LSTM [182]. In [183], a CNN first encoded the image and extracts features, and then an LSTM was used to decode the convolutional features where the weights were learned through attention. A stand-alone self-attention was introduced in [184], which studied whether attention could be used instead of convolution or serve as an interface to support convolutional models. The authors introduced a self-attention layer that can reduce the training parameters while replacing a convolutional layer. The results presented by the authors in this article were obtained by replacing convolutional layers in ResNet with a self-attention block on the imageNet dataset. It is noticeable that the first convolutional layers and the 1×1 convolutional layers remain untouched, and the parameters are reduced by 29% compared to ResNet-50 by introducing a memory block similar to convolution works on a small area at position (i, j).

Recent research such as [185] attempted to overcome the limitations of CNNs due to clothing and pose variation by use of a vision-transformer-based framework built with an attention mechanism.

3.5.4. Global Attention vs Local Attention

Another categorisation of attention is "Global" and "Local" attention. Global Attention is a blend of soft attention, and local attention is a mixture of hard and soft attention. In [186] a local attention mechanism is proposed that essentially improves hard attention and makes it differentiable. Global/Soft attention attends to the whole space [183] on the input sequence, but Local/Hard attention attends to only part of the input space [186].

Humans usually focus on distinctive features in a person's movements and distinguishing characteristics to recognise one another. In other words, their attention is diverted to specific regions in a scene to find salient features. A profound challenge in computer vision is detecting the salient regions of an image [171]. Numerous works in the literature focus on directing the network's attention with saliency maps [171−173]. It has also been used in the past for action recognition problems like [174], which uses selective focus on RGB videos.

A spatial attention mechanism could be used to focus on the important areas of the spatial feature maps. After that, a long short term memory (LSTM) setup could learn temporal gait features. Since the output of the LSTM at each time has a different impact on the performance, we need to assign different weights to each output to control their contribution using a temporal attention mechanism.

For example, [148] used a spatial-temporal attention mechanism to learn gait information from a sequence of silhouette images. Figure 8 shows a high-level overview of this approach which uses an LSTM and attentional interfacing for gait recognition. To summarise, a ResNet extracts spatial feature maps passing them into a spatial attention mechanism that pays attention to the salient regions of each image. Then, an LSTM is used to extract the temporal motion features in order to allow the network to learn when to remember and forget relevant motion information. Since different timesteps of the LSTM have different effects on the network, a temporal attention mechanism is employed to reduce redundancy by giving more weight to the discriminative gait features. This method is evaluated on CASIA-B [187] and OULP from the OU-ISIR Gait database, large population dataset [188] with 4007 subjects to get more view variations. The OULP dataset allows for the determination of the statistically significant performance differences between currently proposed gait features.

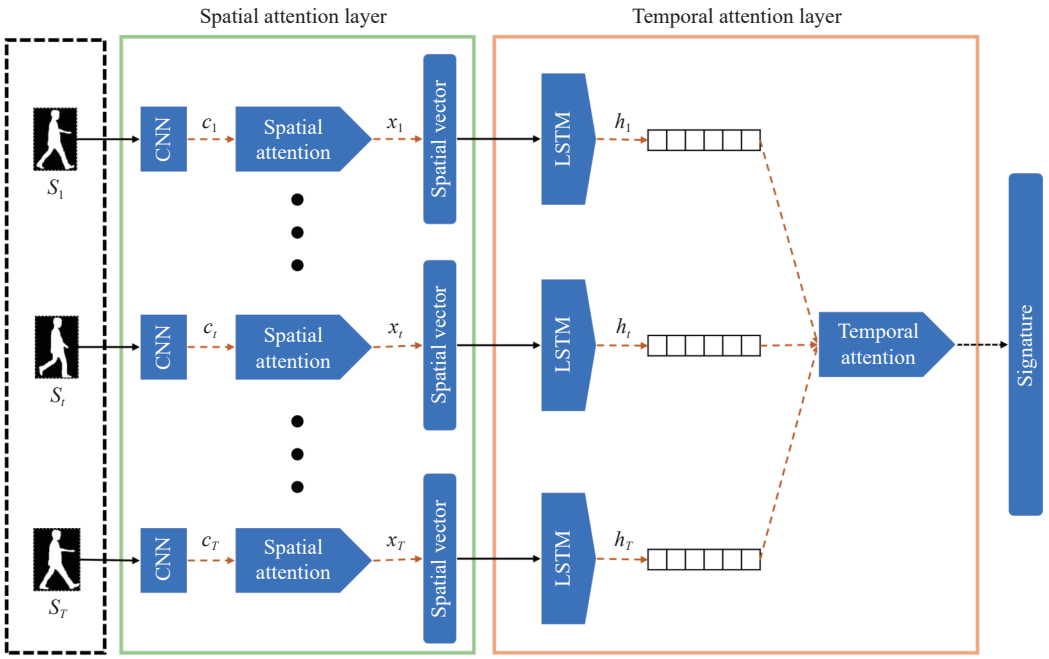

Figure 8. Overview of our spatial-temporal attention approach.

Papers such as [189], as part of the self-supervised learning process, proposed a locality-aware attention mechanism that exploited the locality within individual skeletons of one skeleton sequence to facilitate improved skeleton reconstruction and Gait encoding. Moreover, in [190], adversarial learning was used to extract discriminative Gait features. Additionally, convolutional block embedding locality-aware attention mechanisms were used to extract the rich global information of skeleton data.

4. Conclusion

Gait-based Re-ID is a recent field in pattern recognition that aims at recognising and identifying people by their Gait in unconstrained scenarios typical of video surveillance systems. Unlike classical appearance-based Re-ID approaches, which are used only in short-term situations, gait human Re-ID can be applied to long-term use. Hence, Gait Re-ID offers significant potential for applications in video surveillance, human-robot interaction, and ambient assisted living, among others. In contrast to traditional Gait recognition in controlled setups, Gait Re-ID is a more challenging problem due to person pose, appearance, camera viewpoint, background variety, luminance, and occlusions that are not limited by the scenario and variable.

There are many different techniques to tackle the Gait person re-identification problem. This paper has focused on the most effective methods for gait person Re-ID feature extraction to minimise inter-class and intra-class variations. The purpose of this article is to review the various approaches to Gait-based person re-identification. Several dimensions of the topic have been analysed, such as video acquisition, pose and angle, Gait features extraction and analysis, as well as application scenarios. This article has analysed various approaches in detail, highlighting their strengths and weaknesses. Furthermore, existing Gait person Re-ID datasets and methods using those datasets have been presented to help readers better understand the existing approaches.

Some avenues of gait person Re-ID have not been covered in this paper, including but not limited to, multimodal approaches, which combine sensor data or depth data with the data acquired through surveillance cameras or the fusion of modalities. In addition, graph-based approaches based on pose estimation, body metrics, and body Reconstruction have also been omitted from this paper, which could count as the limitations of this article.

We have paid particular attention to the latest and state-of-the-art feature extraction methods using neural network techniques. We have also discussed some of the main challenges of gait person Re-ID and introduced methods to tackle these challenges using generated data and soft biometrics such as Gait. Since pose and angle have been recognised as one of the most challenging issues in Gait Person Re-ID, we have discussed pose-dependent and pose-independent approaches to the subject.

Based on the methods presented in this article, we have ignored the irregular situation and identified existing gait recognition methods with a main focus on the regular gait cycles. In real-world surveillance, human Gait is almost irregular, which contains arbitrary dynamic characteristics (e.g., duration, speed, and phase) and various viewpoints. This irregularity of Gait poses new challenges for the research community to identify a person and understand

people's movements. Gait recognition in surveillance systems has become a desirable research topic over the past few years because of its across-the-board applications, including social security [191], person Re-ID [120, 192], and analysis of health status [193]. However, due to the complex surveillance environments and diverse human behaviours, human Gait is always irregular.

To be more specific, massive irregular gait sequences contain more than or less than one complete gait cycle, varied lengths, asynchronous paces, different size strides, and inconsistent phases. Moreover, gait appearances are dramatically altered due to camera viewpoints, clothing, and carrying objects. Among these, irregular Gait and change of viewpoints are two significant challenges. The periodic motion cues of the irregular gait sequences are complicated to extract due to the difficulty of precisely detecting the gait cycle. Moreover, the change of viewpoints will seriously magnify the intra-class variations. Additionally, many different subjects with similar gait appearances have unclear inter-class differences, and therefore, are challenging to be accurately identified.

Although existing methods [148] have provided a better avenue for learning gait signatures and improved the gait recognition accuracy to a certain extent, most of them need to detect at least one complete gait cycle precisely and are not robust to the viewpoint changes. Therefore, irregular and specially view-invariant gait recognition still needs particular attention.

On that front, methods with spatial and temporal attention seem most promising. An excellent first line of the study could be the combination of an attentional approach with other variations such as occlusion clothing changes, carry bags and walking speed. As the second line of study, researchers could focus on improving the elements of the existing attentional approach. Since in works such as [148], each architecture component is independent of the others, such components could be replaced relatively easily to test the effect of other state-of-the-art attention mechanisms.

Finally, we identify that the most critical limitation of Gait person Re-ID approaches is the lack of enough training data. Therefore, researchers could look into GANs a means of image generation for supportting Gait person Re-ID tasks.

Funding: This research received no external funding.

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

References

1. Saghafi, M.A.; Hussain, A.; Zaman, H.B.; *et al*. Review of person re-identification techniques. *IET Comput. Vision*, **2014**, *8*: 455−474.
2. Bedagkar-Gala, A.; Shah, S.K. A survey of approaches and trends in person re-identification. *Image Vision Comput.*, **2014**, *32*: 270−286.
3. Lavi, B.; Ullah, I.; Fatan, M.; et al. Survey on reliable deep learning-based person re-identification models: Are we there yet? arXiv preprint arXiv: 2005.00355, 2020. Available online: https://arxiv.org/abs/2005.00355v1 (accessed on 3 December 2022).
4. LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, *Paris*, *France*, *30 May 2010–2 June 2010*; IEEE: Paris, 2010; pp. 253–256. doi: 10.1109/ISCAS.2010.5537907
5. Russakovsky, O.; Deng, J.; Su, H.; *et al*. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision*, **2015**, *115*: 211−252.
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, *Lake Tahoe*, *USA*, *3–6 December 2012*; Curran Associates Inc.: Lake Tahoe, 2012; pp. 1097–1105.
7. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *Proceedings of the 13th European Conference on Computer Vision*, *Zurich*, *Switzerland*, *6–12 September 2014*; Springer: Zurich, 2014; pp. 818–833. doi: 10.1007/978-3-319-10590-1_53
8. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; *et al*. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2010**, *32*: 1627−1645.
9. Redmon, J.; Divvala, S.; Girshick, R.; et al. You only look once: Unified, real-time object detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27–30 June 2016*; IEEE: Las Vegas, 2016; pp. 779–788. doi: 10.1109/CVPR.2016.91
10. Li, J.N.; Liang, X.D.; Shen, S.M.; *et al*. Scale-aware fast R-CNN for pedestrian detection. *IEEE Trans. Multimedia*, **2018**, *20*: 985−996.
11. Girshick, R. Fast R-CNN. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, *Santiago*, *Chile*, *7–13 December 2015*; IEEE: Santiago, 2015; pp. 1440–1448. doi: 10.1109/ICCV.2015.169
12. Liu, W.; Anguelov, D.; Erhan, D.; et al. SSD: Single shot MultiBox detector. In *Proceedings of the 14th European Conference on Computer Vision*, *Amsterdam*, *The Netherlands*, *11–14 October 2016*; Springer: Amsterdam, 2016; pp. 21–37. doi: 10.1007/978-3-319-46448-0_2
13. Dehghan, A.; Assari, S.M.; Shah, M. GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, *Boston*, *USA*, *07–12 June 2015*; IEEE: Boston, 2015; pp. 4091–4099. doi: 10.1109/CVPR.2015.7299036
14. Son, J.; Baek, M.; Cho, M.; et al. Multi-object tracking with quadruplet convolutional neural networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, *Honolulu*, *USA*, *21–26 July 2017*; IEEE: Honolulu, 2017; pp.

3786–3795. doi: 10.1109/CVPR.2017.403

15. Zhu, J.; Yang, H.; Liu, N.; et al. Online multi-object tracking with dual matching attention networks. In *Proceedings of the 15th European Conference on Computer Vision*, *Munich*, *Germany*, *8–14 September 2018*; Springer: Munich, 2018; pp. 379−396. doi: 10.1007/978-3-030-01228-1_23
16. Xu, J.R.; Cao, Y.; Zhang, Z.; et al. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, *Seoul*, *Korea*, *27 October 2019–2 November 2019*; IEEE: Seoul, 2019; pp. 3987−3997. doi: 10.1109/ICCV.2019.00409
17. Zhang, Q.; Cheng, H.J.; Lai, J.H.; et al. DHML: Deep heterogeneous metric learning for VIS-NIR person re-identification. In *Proceedings of the 14th Chinese Conference on Biometric Recognition*, *Zhuzhou*, *China*, *12–13 October 2019*; Springer: Zhuzhou, 2019; pp. 455–465. doi: 10.1007/978-3-030-31456-9_50
18. Martini, M.; Paolanti, M.; Frontoni, E. Open-world person re-identification with RGBD camera in top-view configuration for retail applications. *IEEE Access*, **2020**, *8*: 67756−67765.
19. Liciotti, D.; Paolanti, M.; Frontoni, E.; et al. Person re-identification dataset with RGB-D camera in a top-view configuration. In *International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage*, *Cancun*, *Mexico*, *4 December 2016*; Springer: Cancun, 2016; pp. 1–11. doi: 10.1007/978-3-319-56687-0_1
20. Li, W.; Zhao, R.; Xiao, T.; et al. DeepReiD: Deep filter pairing neural network for person re-identification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *Columbus*, *USA*, *23–28 June 2014*; IEEE: Columbus, 2014; pp. 152–159. doi: 10.1109/CVPR.2014.27
21. Mehta, D.; Sridhar, S.; Sotnychenko, O.; *et al*. Vnect: Real-time 3D human pose estimation with a single RGB camera. *ACM Trans. Graphics*, **2017**, *36*: 44.
22. Zhao, L.M.; Li, X.; Zhuang, Y.T.; et al. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, *Venice*, *Italy*, *22 –29 October 2017*; IEEE: Venice, 2017; pp. 3239–3248. doi: 10.1109/ICCV.2017.349
23. Hirzer, M.; Beleznai, C.; Roth, P.M.; et al. Person re-identification by descriptive and discriminative classification. In *Proceedings of the 17th Scandinavian Conference on Image Analysis*, *Ystad*, *Sweden*, *May 2011*; Springer: Ystad, 2011; pp. 91 –102. doi: 10.1007/978-3-642-21227-7_9
24. Köstinger, M.; Hirzer, M.; Wohlhart, P.; et al. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, *Providence*, *USA*, *16–21 June 2012*; IEEE: Providence, 2012; pp. 2288–2295. doi: 10.1109/CVPR.2012.6247939
25. Liao, S.C.; Hu, Y.; Zhu, X.Y.; et al. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, *Boston*, *USA*, *7–12 June 2015*; IEEE: Boston, 2015; pp. 2197–2206. doi: 10.1109/CVPR.2015.7298832
26. Sun, Y.F.; Zheng, L.; Li, Y.L.; *et al*. Learning part-based convolutional features for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2021**, *43*: 902−917.
27. Li, W.; Zhu, X.T.; Gong, S.G. Harmonious attention network for person re-identification. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City*, *USA*, *18–23 June 2018*; IEEE: Salt Lake City, 2018; pp. 2285–2294. doi: 10.1109/CVPR.2018.00243
28. Chen, D.P.; Li, H.S.; Xiao, T.; et al. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City*, *USA*, *18–23 June 2018*; IEEE: Salt Lake City, 2018; pp. 1169–1178. doi: 10.1109/CVPR.2018.00128
29. Wang, T.Q.; Gong, S.G.; Zhu, X.T.; *et al*. Person re-identification by discriminative selection in video ranking. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2016**, *38*: 2501−2514.
30. McLaughlin, N.; Del Rincon, J.M.; Miller, P. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27–30 June 2016*; IEEE: Las Vegas, 2016; pp. 1325–1334. doi: 10.1109/CVPR.2016.148
31. Zhu, X.K.; Jing, X.Y.; You, X.E.; *et al*. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. *IEEE Trans. Image Process.*, **2018**, *27*: 5683−5695.
32. Zheng, L.; Bie, Z.; Sun, Y.F.; et al. MARS: A video benchmark for large-scale person re-identification. In *Proceedings of the 14th European Conference on Computer Vision*, *Amsterdam*, *The Netherlands*, *11–14 October 2016*; Springer: Amsterdam, 2016; pp. 868–884. doi: 10.1007/978-3-319-46466-4_52
33. Liu, K.; Ma, B.P.; Zhang, W.; et al. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, *Santiago*, *Chile*, *7–13 December 2015*; IEEE: Santiago, 2015; pp. 3810–3818. doi: 10.1109/ICCV.2015.434
34. You, J.J.; Wu, A.C.; Li, X.; et al. Top-push video-based person re-identification. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27–30 June 2016*; IEEE: Las Vegas, 2016; pp. 1345 –1353. doi: 10.1109/CVPR.2016.150
35. Subramaniam, A.; Nambiar, A.; Mittal, A. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, *Seoul*, *Korea*, *27 October 2019–2 November 2019*; IEEE: Seoul, 2019; pp. 562–572. doi: 10.1109/ICCV.2019.00065
36. Hirzer, M.; Roth, P.M.; Köstinger, M.; et al. Relaxed pairwise learned metric for person re-identification. In *Proceedings of the 12th European Conference on Computer Vision*, *Florence*, *Italy*, *7–13 October 2012*; Springer: Florence, 2012; pp. 780–793. doi: 10.1007/978-3-642-33783-3_56
37. Wu, S.X.; Chen, Y.C.; Li, X.; et al. An enhanced deep feature representation for person re-identification. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, *Lake Placid*, *USA*, *7–10 March 2016*; IEEE: Lake Placid, 2016; pp. 1–8. doi: 10.1109/WACV.2016.7477681
38. Xiong, F.; Gou, M.R.; Camps, O.; et al. Person re-identification using kernel-based metric learning methods. In *Proceedings of the 13th European Conference on Computer Vision*, *Zurich*, *Switzerland*, *6–12 September 2014*; Springer: Zurich, 2014; pp. 1–16. doi: 10.1007/978-3-319-10584-0_1
39. Zheng, W.S.; Gong, S.G.; Xiang, T. Towards open-world person re-identification by one-shot group-based verification. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2016**, *38*: 591−606.
40. Gray, D.; Tao, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the 10th*

European Conference on Computer Vision, *Marseille*, *France*, *12–18 October 2008*; Springer: Marseille, 2008; pp. 262–275. doi: 10.1007/978-3-540-88682-2_21

41. Kuo, C.H.; Khamis, S.; Shet, V. Person re-identification using semantic color names and rankboost. In *2013 IEEE workshop on applications of computer vision (WACV)*, *Clearwater Beach*, *USA*, *15–17 January 2013*; IEEE: Clearwater Beach, 2013; pp. 281–287. doi: 10.1109/WACV.2013.6475030
42. Cheng, D.; Gong, Y.H.; Zhou, S.P.; et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27–30 June 2016*; IEEE: Las Vegas, 2016; pp. 1335–1344. doi: 10.1109/CVPR.2016.149
43. Xiao, T.; Li, H.S.; Ouyang, W.L.; et al. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27–30 June 2016*; IEEE: Las Vegas, 2016; pp. 1249–1258. doi: 10.1109/CVPR.2016.140
44. Pedagadi, S.; Orwell, J.; Velastin, S.; et al. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, *Portland*, *USA*, *23–28 June 2013*; IEEE: Portland, 2013; pp. 3318–3325. doi: 10.1109/CVPR.2013.426
45. Sugiyama, M. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, *Pittsburgh*, *USA*, *25 –29 June 2006*; ACM: Pittsburgh, 2006; pp. 905 –912. doi: 10.1145/1143844.1143958
46. Weinberger, K.Q.; Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, **2009**, *10*: 207−244.
47. Zhang, L.; Xiang, T.; Gong, S.G. Learning a discriminative null space for person re-identification. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27 –30 June 2016*; IEEE: Las Vegas, 2016; pp. 1239–1248. doi: 10.1109/CVPR.2016.139
48. Lu, M. Ranked: The world's most surveilled cities. Oct 2022. Available online: https://www.visualcapitalist.com/ranked-the-worlds-most-surveilled-cities/#:~:text=IHS%20Markit%20estimates%20that%20as,billion%20sureillance%20cameras%20installed%20worldwide (accessed on 3 December 2022).
49. BBC.co.uk. CCTV: Too many cameras useless, warns surveillance watchdog tony porter. Jan 2015. Available online: https://www.timefortruth.eu/cctv-too-many-cameras-useless-warns-surveillance-watchdog-tony-porter/ (accessed on 4 December 2022).
50. Cong, D.N.T.; Achard, C.; Khoudour, L.; et al. Video sequences association for people re-identification across multiple non-overlapping cameras. In *Proceedings of the 15th International Conference on Image Analysis and Processing*, *Vietri sul Mare*, *Italy*, *8–11 September 2009*; Springer: Vietri sul Mare, 2009; pp. 179–189. doi: 10.1007/978-3-642-04146-4_21
51. Guermazi, R.; Hammami, M.; Hamadou, A.B. Violent web images classification based on MPEG7 color descriptors. In *2009 IEEE International Conference on Systems*, *Man and Cybernetics*, *San Antonio*, *USA*, *11–14 October 2009*; IEEE: San Antonio, 2009; pp. 3106–3111. doi: 10.1109/ICSMC.2009.5346149
52. Zheng, L.; Shen, L.Y.; Tian, L.; et al. Scalable person re-identification: A benchmark. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, *Santiago*, *Chile*, *7–13 December 2015*; IEEE: Santiago, 2015; pp. 1116–1124. doi: 10.1109/ICCV.2015.133
53. Lyons, M.J.; Budynek, J.; Akamatsu, S. Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1999**, *21*: 1357−1362.
54. Álvarez-Aparicio, C.; Guerrero-Higueras, Á.M.; González-Santamarta, M.Á.; *et al*. Biometric recognition through gait analysis. *Sci. Rep.*, **2022**, *12*: 14530.
55. Alobaidi, H.; Clarke, N.; Li, F.D.; *et al*. Real-world smartphone-based gait recognition. *Comput. Secur.*, **2022**, *113*: 102557.
56. Maloney, L.T.; Wandell, B.A. Color constancy: A method for recovering surface spectral reflectance. *J. Opt. Soc. Am. A*, **1986**, *3*: 29−33.
57. Qian, X.L.; Wang, W.X.; Zhang, L.; et al. Long-term cloth-changing person re-identification. In *Proceedings of the 15th Asian Conference on Computer Vision*, *Kyoto*, *Japan*, *30 November –4 December 2020*; Springer: Kyoto, 2020; pp. 71 –88. doi: 10.1007/978-3-030-69535-4_5
58. Hirzer, M.; Beleznai, C.; Roth, P.M.; et al. Person re-identification by descriptive and discriminative classification. In *Proceedings of the 17th Scandinavian Conference on Image Analysis*, *Ystad*, *Sweden*, *May 2011*; Springer: Ystad, 2011; pp. 91 –102. doi: 10.1007/978-3-642-21227-7_9
59. Li, M.X.; Zhu, X.T.; Gong, S.G. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the 15th European Conference on Computer Vision*, *Munich*, *Germany*, *8–14*, *September 2018*; Springer: Munich, 2018; pp. 772–788. doi: 10.1007/978-3-030-01225-0_45
60. Zheng, Z.D.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, *Venice*, *Italy*, *22–29 October 2017*; IEEE: Venice, 2017; pp. 3774–3782. doi: 10.1109/ICCV.2017.405
61. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; et al. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, *Montreal Canada*, *8 –13 December 2014*; MIT Press: Montreal, 2014; pp. 2672–2680.
62. Wei, L.H.; Zhang, S.L.; Gao, W.; et al. Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City*, *USA*, *18–23 June 2018*; IEEE: Salt Lake City, 2018; pp. 79–88. doi: 10.1109/CVPR.2018.00016
63. Dai, C.Q.; Peng, C.; Chen, M. Selective transfer cycle GAN for unsupervised person re-identification. *Multimed. Tools Appl.*, **2020**, *79*: 12597−12613.
64. Qian, X.L.; Fu, Y.W.; Xiang, T.; et al. Pose-normalized image generation for person re-identification. In *Proceedings of the 15th European Conference on Computer Vision*, *Munich*, *Germany*, *8–14 September 2018*; Springer: Munich, 2018; pp. 661–678. doi: 10.1007/978-3-030-01240-3_40
65. Fan, H.H.; Zheng, L.; Yan, C.G.; *et al*. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Trans. Multimedia Comput., Commun., Appl.*, **2018**, *14*: 83.
66. Lin, Y.T.; Dong, X.Y.; Zheng, L.; *et al*. A bottom-up clustering approach to unsupervised person re-identification. *Proc. AAAI Conf. Artif. Intell.*, **2019**, *33*: 8738−8745.
67. Ding, Y.H.; Fan, H.H.; Xu, M.L.; *et al*. Adaptive exploration for unsupervised person re-identification. *ACM Trans. Multimedia*

Comput., Commun., Appl., **2020**, *16*: 3.
68. Pala, F.; Satta, R.; Fumera, G.; *et al*. Multimodal person reidentification using rgb-d cameras. *IEEE Trans. Circuits Syst. Video Technol.*, **2016**, *26*: 788−799.
69. Kniaz, V.V.; Knyaz, V.A.; Hladůvka, J.; et al. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision*, *Munich*, *Germany*, *8–14 September 2018*; Springer: Munich, 2018; pp. 606–624. doi: 10.1007/978-3-030-11024-6_46
70. Elsken, T.; Metzen, J.H.; Hutter, F. Correction to: Neural architecture search. In *Automated Machine Learning*; Hutter, F.; Kotthoff, L.; Vanschoren, J., Eds.; Springer: Cham, 2019; p. C1. doi: 10.1007/978-3-030-05318-5_11
71. Zhou, K.Y.; Yang, Y.X.; Cavallaro, A.; Xiang, T. Learning generalisable Omni-scale representations for person re-identification. arXiv preprint arXiv: 1910.06827, 2019. Available online: https://arxiv.org/abs/1910.06827 (accessed on 5 December 2022).
72. Ross, A.; Shah, J.; Jain, A.K. From template to image: Reconstructing fingerprints from minutiae points. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2007**, *29*: 544−560.
73. Dantcheva, A.; Dugelay, J.L.; Elia, P. Soft biometrics systems: Reliability and asymptotic bounds. In *2010 Fourth IEEE International Conference on Biometrics: Theory*, *Applications and Systems (BTAS)*, *Washington*, *USA*, *27–29 September 2010*; IEEE: Washington, 2010; pp. 1–6. doi: 10.1109/BTAS.2010.5634534
74. Nixon, M.S.; Tan, T.; Chellappa, R. *Human Identification Based on Gait*; Springer Science & Business Media, 2010; pp. 4. Available online: https://link.springer.com/referenceworkentry/10.1007/978-0-387-31439-6_375 (accessed on 5 December 2022).
75. Cordina, C. Face blindness. MMSA, 2020. Available online: https://www.um.edu.mt/library/oar/handle/123456789/52399 (accessed on 5 December 2022).
76. Bate, S.; Bennetts, R.J. The rehabilitation of face recognition impairments: A critical review and future directions. *Front. Hum. Neurosci.*, **2014**, *8*: 491.
77. DeGutis, J.M.; Chiu, C.; Grosso, M.E.; *et al*. Face processing improvements in prosopagnosia: Successes and failures over the last 50 years. *Front. Hum. Neurosci.*, **2014**, *8*: 561.
78. Di Biase, L.; Di Santo, A.; Caminiti, M.L.; *et al*. Gait analysis in Parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, **2020**, *20*: 3529.
79. White, H.; Augsburger, S. Gait evaluation for patients with cerebral palsy. In *Orthopedic Care of Patients with Cerebral Palsy*; Nowicki, P.D., Ed., Springer: Cham, 2020; pp. 51–76. doi: 10.1007/978-3-030-46574-2_4
80. Li, M.X.; Tian, S.S.; Sun, L.L.; *et al*. Gait analysis for post-stroke hemiparetic patient by multi-features fusion method. *Sensors*, **2019**, *19*: 1737.
81. Whittle, M.W. Clinical gait analysis: A review. *Human Mov. Sci.*, **1996**, *15*: 369−387.
82. Zijlstra, W.; Hof, A.L. Assessment of spatio-temporal gait parameters from trunk accelerations during human walking. *Gait Posture*, **2003**, *18*: 1−10.
83. Roy, A.; Sural, S.; Mukherjee, J. A hierarchical method combining gait and phase of motion with spatiotemporal model for person re-identification. *Pattern Recognit. Lett.*, **2012**, *33*: 1891−1901.
84. Ariyanto, G.; Nixon, M.S. Model-based 3D gait biometrics. In *2011 International Joint Conference on Biometrics (IJCB)*, *Washington*, *USA*, *11–13 October 2011*; IEEE: Washington, 2011; pp. 1–7. doi: 10.1109/IJCB.2011.6117582
85. Goffredo, M.; Bouchrika, I.; Carter, J.N.; *et al*. Self-calibrating view-invariant gait biometrics. *IEEE Trans. Syst., Man, Cybern., Part B (Cybern.)*, **2010**, *40*: 997−1008.
86. Lombardi, S.; Nishino, K.; Makihara, Y.; et al. Two-point gait: Decoupling gait from body shape. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, *Sydney*, *Australia*, *1–8 December 2013*; IEEE: Sydney, 2013; pp. 1041−1048. doi: 10.1109/ICCV.2013.133
87. Lu, H.P.; Plataniotis, K.N.; Venetsanopoulos, A.N. A layered deformable model for gait analysis. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, *Southampton*, *10–12 April 2006*; IEEE: Southampton, 2006; pp. 249–254. doi: 10.1109/FGR.2006.11
88. Khamsemanan, N.; Nattee, C.; Jianwattanapaisarn, N. Human identification from freestyle walks using posture-based gait feature. *IEEE Trans. Inf. Forensics Secur.*, **2018**, *13*: 119−128.
89. Bouchrika, I.; Carter, J.N.; Nixon, M.S. Towards automated visual surveillance using gait for identity recognition and tracking across multiple non-intersecting cameras. *Multimed. Tools Appl.*, **2016**, *75*: 1201−1221.
90. Yandell, M.B.; Quinlivan, B.T.; Popov, D.; *et al*. Physical interface dynamics alter how robotic exosuits augment human movement: Implications for optimizing wearable assistive devices. *J. Neuroeng. Rehabil*, **2017**, *18*(40).
91. Muybridge, E. Animal locomotion: An Electro-photographic investigation of consecutives phases of animal movement: Prospectus and catalogue of plates.. Males (nude). I., 1969. Available online: https://www.metmuseum.org/art/collection/search/266429 (accessed on 5 December 2022).
92. Dagognet, F. Lanimal selon condillac, introd. a eb de condillac, traité des animaux, Vrin, j.; Paris; pp. 7−131, 1987. Available online: https://philpapers.org/rec/DAGLSC (accessed on 5 December 2022).
93. Balazia, M.; Sojka, P. Gait recognition from motion capture data. *ACM Trans. Multimedia Comput., Commun., Appl.*, **2018**, *14*: 22.
94. Josiński, H.; Michalczuk, A.; Kostrzewa, D.; et al. Heuristic method of feature selection for person re-identification based on gait motion capture data. In *Proceedings of the 6th Asian Conference on Intelligent Information and Database Systems*, *Bangkok*, *Thailand*, *7–9 April 2014*; Springer: Bangkok, 2014; pp. 585–594. doi: 10.1007/978-3-319-05458-2_60
95. Nambiar, A.M.; Bernardino, A.; Nascimento, J.C.; et al. Towards view-point invariant person re-identification via fusion of anthropometric and gait features from kinect measurements. In *Proceedings of the 12th International Joint Conference on Computer Vision*, *Imaging and Computer Graphics Theory and Applications*, *Porto*, *Portugal*, *27 February–1 March 2017*; VISAPP: Porto, 2017; pp. 108–119.
96. Hofmann, M.; Geiger, J.; Bachmann, S.; *et al*. The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits. *J. Visual Commun. Image Representat.*, **2014**, *25*: 195−206.
97. Bedagkar-Gala, A.; Shah, S.K. Gait-assisted person re-identification in wide area surveillance. In *Asian Conference on Computer Vision*, *Singapore*, *1–2 November 2014*; Springer: Singapore, 2014; pp. 633–649. doi: 10.1007/978-3-319-16634-6_46
98. Wei, L.; Tian, Y.H.; Wang, Y.W.; et al. Swiss-system based cascade ranking for gait-based person re-identification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, *Austin Texas*, *USA*, *25–30 January 2015*; AAAI Press: Austin Texas, 2015; pp. 1882–1888.
99. Liu, Z.; Zhang, Z.X.; Wu, Q.; *et al*. Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, **2015**, *168*:

1144–1156.
100. Iwashita, Y.; Baba, R.; Ogawara, K.; et al. Person identification from spatio-temporal 3D gait. In *2010 International Conference on Emerging Security Technologies*, *Canterbury*, *UK*, *6 –7 September 2010*; IEEE: Canterbury, 2010; pp. 30 –35. doi: 10.1109/EST.2010.19
101. Nambiar, A.; Bernardino, A.; Nascimento, J.C.; et al. Context-aware person re-identification in the wild via fusion of gait and anthropometric features. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, *Washington*, *USA*, *30 May 2017–3 June 2017*; IEEE: Washington, 2017; pp. 973–980. doi: 10.1109/FG.2017.121
102. John, V.; Englebienne, G.; Krose, B. Person re-identification using height-based gait in colour depth camera. In *2013 IEEE International Conference on Image Processing*, *Melbourne*, *Australia*, *15–18 September 2013*; IEEE: Melbourne, 2013; pp. 3345–3349. doi: 10.1109/ICIP.2013.6738689
103. Nambiar, A.; Nascimento, J.C.; Bernardino, A.; et al. Person re-identification in frontal gait sequences via histogram of optic flow energy image. In *Proceedings of the 17th International Conference on Advanced Concepts for Intelligent Vision Systems*, *Lecce*, *Italy*, *24–27 October 2016*; Springer: Lecce, 2016; pp. 250–262. doi: 10.1007/978-3-319-48680-2_23
104. Chattopadhyay, P.; Sural, S.; Mukherjee, J. Information fusion from multiple cameras for gait-based re-identification and recognition. *IET Image Process.*, **2015**, *9*: 969–976.
105. Kawai, R.; Makihara, Y.; Hua, C.S.; et al. Person re-identification using view-dependent score-level fusion of gait and color features. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, *Tsukuba*, *Japan*, *11–15 November 2012*; IEEE: Tsukuba, 2012; pp. 2694–2697.
106. Wang, T.Q.; Gong, S.G.; Zhu, X.T.; et al. Person re-identification by video ranking. In *Proceedings of the 13th European Conference on Computer Vision*, *Zurich*, *Switzerland*, *6–12 September 2014*; Springer: Zurich, 2014; pp. 688–703. doi: 10.1007/978-3-319-10593-2_45
107. Zhang, S.X.; Wang, Y.Z.; Chai, T.R.; et al. Realgait: Gait recognition for person re-identification. arXiv preprint arXiv: 2201.04806, 2022. Available online: https://arxiv.org/abs/2201.04806 (accessed on 5 December 2022).
108. Balazia, M.; Sojka, P. You are how you walk: Uncooperative MoCap gait identification for video surveillance with incomplete and noisy data. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, *Denver*, *USA*, *1–4 October 2017*; IEEE: Denver, 2017; pp. 208–215. doi: 10.1109/BTAS.2017.8272700
109. Nambiar, A.M.; Bernardino, A.; Nascimento, J.C. Cross-context analysis for long-term view-point invariant person re-identification via soft-biometrics using depth sensor. In *Proceedings of the 13th International Joint Conference on Computer Vision*, *Imaging and Computer Graphics Theory and Applications*, *Funchal*, *Portugal*, *27–29 January 2018*; VISAPP: Funchal, 2018; pp. 105–113.
110. Stöckel, T.; Jacksteit, R.; Behrens, M.; *et al*. The mental representation of the human gait in young and older adults. *Front. Psychol.*, **2015**, *6*: 943.
111. Burnfield, M. Gait analysis: Normal and pathological function. *J. Sports Sci. Med*. **2010**, *9*, 353.
112. Gabel, M.; Gilad-Bachrach, R.; Renshaw, E.; et al. Full body gait analysis with kinect. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, *San Diego*, *USA*, *28 August 2012–1 September 2012*; IEEE: San Diego, 2012; pp. 1964–1967. doi: 10.1109/EMBC.2012.6346340
113. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2001**, *23*: 257–267.
114. Balazia, M.; Plataniotis, K.N. Human gait recognition from motion capture data in signature poses. *IET Biom.*, **2017**, *6*: 129–137.
115. Han, J.; Bhanu, B. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2006**, *28*: 316–322.
116. Lam, T.H.W.; Cheung, K.H.; Liu, J.N.K. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognit.*, **2011**, *44*: 973–987.
117. Castro, F.M.; Marín-Jimenez, M.J.; Medina-Carnicer, R. Pyramidal fisher motion for multiview gait recognition. In *2014 22nd International Conference on Pattern Recognition*, *Stockholm*, *Sweden*, *24 –28 August 2014*; IEEE: Stockholm, 2014; pp. 1692–1697. doi: 10.1109/ICPR.2014.298
118. Tang, J.; Luo, J.; Tjahjadi, T.; *et al*. Robust arbitrary-view gait recognition based on 3D partial similarity matching. *IEEE Trans. Image Process.*, **2017**, *26*: 7–22.
119. Connie, T.; Goh, M.K.O.; Teoh, A.B.J. A grassmannian approach to address view change problem in gait recognition. *IEEE Trans. Cybern.*, **2017**, *47*: 1395–1408.
120. Zhang, C.; Liu, W.; Ma, H.D.; et al. Siamese neural network based gait recognition for human identification. In *2016 IEEE international conference on acoustics*, *speech and signal processing (ICASSP)*, *Shanghai*, *China*, *20–25 March 2016*; IEEE: Shanghai, 2016; pp. 2832–2836. doi: 10.1109/ICASSP.2016.7472194
121. Wu, Z.F.; Huang, Y.Z.; Wang, L.; *et al*. A comprehensive study on cross-view gait based human identification with deep CNNs. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2017**, *39*: 209–226.
122. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *San Diego*, *USA*, *20–25 June 2005*; IEEE: San Diego, 2005; pp. 539–546. doi: 10.1109/CVPR.2005.202
123. Shen, C.; Jin, Z.M.; Zhao, Y.R.; et al. Deep Siamese network with multi-level similarity perception for person re-identification. In *Proceedings of the 25th ACM international conference on Multimedia*, *Mountain*, *USA*, *23–27 October 2017*; ACM: Mountain, 2017; pp. 1942–1950. doi: 10.1145/3123266.3123452
124. Wu, L.; Wang, Y.; Gao, J.B.; *et al*. Where-and-when to look: Deep Siamese attention networks for video-based person re-identification. *IEEE Trans. Multimedia*, **2019**, *21*: 1412–1424.
125. Zheng, M.; Karanam, S.; Wu, Z.Y.; et al. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Long Beach*, *USA*, *15–20 June 2019*; IEEE: Long Beach, 2019; pp. 5728–5737. doi: 10.1109/CVPR.2019.00588
126. Bromley, J.; Bentz, J.W.; Bottou, L.; *et al*. Signature verification using a "Siamese" time delay neural network. *Int. J. Pattern Recognit. Artif. Intell.*, **1993**, *7*: 669–688.
127. Lu, T.; Zhou, Q.; Fang, W.H.; *et al*. Discriminative metric learning for face verification using enhanced Siamese neural network. *Multimed. Tools Appl.*, **2021**, *80*: 8563–5880.
128. Liu, W.; Zhang, C.; Ma, H.D.; *et al*. Learning efficient spatial-temporal gait features with deep learning for human identification. *Neuroinformatics*, **2018**, *16*: 457–471.

129. Li, S.Q.; Liu, W.; Ma, H.D.; et al. Beyond view transformation: Cycle-consistent global and partial perception gan for view-invariant gait recognition. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, *San Diego*, *USA*, *23–27 July 2018*; IEEE: San Diego, 2018; pp. 1–6. doi: 10.1109/ICME.2018.8486484
130. Carley, C.; Ristani, E.; Tomasi, C. Person re-identification from gait using an autocorrelation network. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, *Long Beach*, *USA*, *16–17 June 2019*; IEEE: Long Beach, 2019; pp. 2345–2353. doi: 10.1109/CVPRW.2019.00288
131. He, Y.W.; Zhang, J.P.; Shan, H.M.; *et al*. Multi-task GANs for view-specific feature learning in gait recognition. *IEEE Trans. Inf. Forensics Secur.*, **2019**, *14*: 102−113.
132. Wang, X.H.; Feng, S.L.; Yan, W.Q. Human gait recognition based on self-adaptive hidden Markov model. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **2021**, *18*: 963−972.
133. Xu, Y.J.; Han, Y.H.; Hong, R.C.; *et al*. Sequential video VLAD: Training the aggregation locally and temporally. *IEEE Trans. Image Process.*, **2018**, *27*: 4933−4944.
134. Zhao, S.C.; Liu, Y.B.; Han, Y.H.; *et al*. Pooling the convolutional layers in deep convnets for video action recognition. *IEEE Trans. Circuits Syst. Video Technol.*, **2018**, *28*: 1839−1849.
135. Nguyen, P.; Han, B.; Liu, T.; et al. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City*, *USA*, *18–23 June 2018*; IEEE: Salt Lake City, 2018; pp. 6752–6761. doi: 10.1109/CVPR.2018.00706
136. Chao, H.Q.; He, Y.W.; Zhang, J.P.; *et al*. GaitSet: Regarding gait as a set for cross-view gait recognition. *Proc. AAAI Conf. Artif. Intell.*, **2019**, *33*: 8126−8133.
137. Li, S.Q.; Liu, W.; Ma, H.D. Attentive spatial–temporal summary networks for feature learning in irregular gait recognition. *IEEE Trans. Multimedia*, **2019**, *21*: 2361−2375.
138. Hu, B.Z.; Gao, Y.; Guan, Y.; et al. Robust cross-view gait identification with evidence: A discriminant gait GAN (DiGGAN) approach on 10000 people. arXiv preprint arXiv: 1811.10493, 2018. Available online: https://arxiv.org/abs/1811.10493 (accessed on 5 December 2022).
139. Wang, Y.Y.; Song, C.F.; Huang, Y.; *et al*. Learning view invariant gait features with two-stream GAN. *Neurocomputing*, **2019**, *339*: 245−254.
140. Zhang, Z.Y.; Tran, L.; Liu, F.; *et al*. On learning disentangled representations for gait recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2022**, *44*: 345−360.
141. Liao, R.J.; Yu, S.Q.; An, W.Z.; *et al*. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognit.*, **2020**, *98*: 107069.
142. Burges, C.J.C. *Dimension Reduction: A Guided Tour*; Now Publishers Inc: Boston, 2010.
143. Avraham, T.; Lindenbaum, M. Learning appearance transfer for person re-identification. In *Person Re-Identification*; Gong, S.G.; Cristani, M.; Yan, S.C., et al, Eds.; Springer: London, 2014; pp. 231–246. doi: 10.1007/978-1-4471-6296-4_11
144. Castro, F.M.; Marín-Jiménez, M.J.; Guil, N.; et al. Evaluation of CNN architectures for gait recognition based on optical flow maps. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, *Darmstadt*, *Germany*, *20–22 September 2017*; IEEE: Darmstadt, 2017; pp. 1–5. doi: 10.23919/BIOSIG.2017.8053503
145. Castro, F.M.; Marín-Jiménez, M.J.; Guil, N. Multimodal features fusion for gait, gender and shoes recognition. *Mach. Vision Appl.*, **2016**, *27*: 1213−1228.
146. He, K.M.; Zhang, X.Y.; Ren, S.Q.; et al. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27–30 June 2016*; IEEE: Las Vegas, 2016; pp. 770–778. doi: 10.1109/CVPR.2016.90
147. Wang, H.; Kläser, A.; Schmid, C.; et al. Action recognition by dense trajectories. In *CVPR 2011*, *Colorado Springs*, *USA*, *20–25 June 2011*; IEEE: Colorado Springs, 2011; pp. 3169–3176. doi: 10.1109/CVPR.2011.5995407
148. Rahi, B. View-Invariant Gait Person Re-Identification with Spatial and Temporal Attention. Ph.D. Thesis, Brunel University London, 2021. Available online: https://bura.brunel.ac.uk/handle/2438/24380 (accessed on 5 December 2022).
149. Jain, M.; Jégou, H.; Bouthemy, P. Better exploiting motion for better action recognition. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, *Portland*, *USA*, *23–28 June 2013*; IEEE: Portland, 2013; pp. 2555–2562. doi: 10.1109/CVPR.2013.330
150. Perronnin, F.; Larlus, D. Fisher vectors meet neural networks: A hybrid classification architecture. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, *Boston*, *USA*, *7–12 June 2015*; IEEE: Boston, 2015; pp. 3743–3752. doi: 10.1109/CVPR.2015.7298998
151. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision*, *Heraklion*, *Crete*, *Greece*, *5 –11 September 2010*; Springer: Heraklion, 2010; pp. 143–156. doi: 10.1007/978-3-642-15561-1_11
152. Jaakkola, T.S.; Haussler, D. Exploiting generative models in discriminative classifiers. In *Proceedings of the 11th International Conference on Neural Information Processing Systems*, *Denver*, *USA*, *1 –3 December 1998*; MIT Press: Denver, 1999; pp. 487–493.
153. Castro, F.M.; Marín-Jiménez, M.J.; Guil, N.; et al. Automatic learning of gait signatures for people identification. In *Proceedings of the 14th International Work-Conference on Artificial Neural Networks*, *Cadiz*, *Spain*, *14–16 June 2017*; Springer: Cadiz, 2017; pp. 257–270. doi: 10.1007/978-3-319-59147-6_23
154. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, *Montreal*, *Canada*, *8–13 December 2014*; MIT Press: Montreal, 2014; pp. 568–576.
155. Zhang, B.W.; Wang, L.M.; Wang, Z.; et al. Real-time action recognition with enhanced motion vector CNNs. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27–30 June 2016*; IEEE: Las Vegas, 2016; pp. 2718–2726. doi: 10.1109/CVPR.2016.297
156. Piergiovanni, A.J.; Ryoo, M.S. Representation flow for action recognition. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Long Beach*, *USA*, *15–20 June 2019*; IEEE: Long Beach, 2019; pp. 9937–9945. doi: 10.1109/CVPR.2019.01018
157. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, *Halmstad*, *Sweden*, *29 June–2 July 2003*; Springer: Halmstad, 2003; pp. 363–370. doi: 10.1007/3-540-45103-

X_50
158. Horn, B.K.P.; Schunck, B.G. Determining optical flow. In *Proceedings of SPIE 0281, Techniques and Applications of Image Understanding*, *Washington*, *USA*, *12 November1981*; SPIE: Washington, 1981; pp. 319–331. doi: 10.1117/12.965761
159. Lucas, B.D.; Kanade, T. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, *Vancouver*, *Canada*, *24–28 August 1981*; Morgan Kaufmann Publishers Inc.: Vancouver, 1981; pp. 674–679.
160. Brox, T.; Bruhn, A.; Papenberg, N.; et al. High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the 8th European Conference on Computer Vision*, *Prague*, *Czech Republic*, *11–14 May 2004*; Springer: Prague, 2004; pp. 25–36. doi: 10.1007/978-3-540-24673-2_3
161. Papenberg, N.; Bruhn, A.; Brox, T.; *et al*. Highly accurate optic flow computation with theoretically justified warping. *Int. J. Comput. Vision*, **2006**, *67*: 141−158.
162. Brox, T.; Malik, J. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2010**, *33*: 500−513.
163. Hui, T.W.; Chung, R. Determining motion directly from normal flows upon the use of a spherical eye platform. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, *Portland*, *USA*, *23–28 June 2013*; IEEE: Portland, 2013; pp. 2267–2274. doi: 10.1109/CVPR.2013.294
164. Hui, T.W.; Chung, R. Determining shape and motion from monocular camera: A direct approach using normal flows. *Pattern Recognit.*, **2015**, *48*: 422−437.
165. Dosovitskiy, A.; Fischer, P.; Ilg, E.; et al. FlowNet: Learning optical flow with convolutional networks. In P*roceedings of the 2015 IEEE International Conference on Computer Vision*, *Santiago*, *Chile*, *7–13 December 2015*; IEEE: Santiago, 2015; pp. 2758–2766. doi: 10.1109/ICCV.2015.316
166. Ilg, E.; Mayer, N.; Saikia, T.; et al. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, *Honolulu*, *USA*, *21–26 July 2017*; IEEE: Honolulu, 2017; pp. 1647–1655. doi: 10.1109/CVPR.2017.179
167. Hui, T.W.; Tang, X.O.; Loy, C.C. LiteFlowNet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City*, *USA*, *18–23 June 2018*; IEEE: Salt Lake City, 2018; pp. 8981–8989. doi: 10.1109/CVPR.2018.00936
168. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, *Las Vegas*, *USA*, *27–30 June 2016*; IEEE: Las Vegas, 2016; pp. 1933–1941. doi: 10.1109/CVPR.2016.213
169. Tran, D.; Bourdev, L.; Fergus, R.; et al. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, *Santiago*, *Chile*, *7–13 December 2015*; IEEE: Santiago, 2015; pp. 4489–4497. doi: 10.1109/ICCV.2015.510
170. Yu, S.Q.; Chen, H.F.; Wang, Q.; *et al*. Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing*, **2017**, *239*: 81−93.
171. Goferman, S.; Zelnik-Manor, L.; Tal, A. Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2012**, *34*: 1915−1926.
172. Ba, J.; Mnih, V.; Kavukcuoglu, K. Multiple object recognition with visual attention. In *Proceedings of the 3rd International Conference on Learning Representations*, *San Diego*, *USA*, *7–9 May 2015*; ICLR: San Diego, 2014.
173. Bazzani, L.; Larochelle, H.; Torresani, L. Recurrent mixture density network for spatiotemporal visual attention. In *Proceedings of the 5th International Conference on Learning Representations*, *Toulon*, *France*, *24–26 April 2017*; ICLR: Toulon, 2017.
174. Sharma, S.; Kiros, R.; Salakhutdinov, R. Action recognition using visual attention. arXiv preprint arXiv: 1511.04119, 2015. Available online: https://arxiv.org/abs/1511.04119 (accessed on 5 December 2022).
175. Sepas-Moghaddam, A.; Etemad, A. View-invariant gait recognition with attentive recurrent learning of partial representations.. *IEEE Trans. Biom., Behav., Identity Sci*, **2021**, *3*: 124−137.
176. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, *San Diego*, *USA*, *7–9 May 2015*; ICLR: San Diego, 2015.
177. Vinyals, O.; Kaiser, L.; Koo, T.; et al. Grammar as a foreign language. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, *Montreal*, *Canada*, *7–12 December 2015*; MIT Press: Montreal, 2015; pp. 2773–2781.
178. Chan, W.; Jaitly, N.; Le, Q.V.; et al. Listen, attend and spell. arXiv preprint arXiv: 1508.01211, 2015. Available online: https://arxiv.org/abs/1508.01211 (accessed on 5 December 2022).
179. Guan, Q.J.; Huang, Y.P.; Zhong, Z.; et al. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv: 1801.09927, 2018. Available online: https://arxiv.org/abs/1801.09927 (accessed on 5 December 2022).
180. Yao, L.; Poblenz, E.; Dagunts, D.; et al. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv: 1710.10501, 2017. Available online: https://arxiv.org/abs/1710.10501 (accessed on 5 December 2022).
181. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Salt Lake City*, *USA*, *18–23 June 2018*; IEEE: Salt Lake City, 2018; pp. 7132–7141. doi: 10.1109/CVPR.2018.00745
182. Cheng, J.P.; Dong, L.; Lapata, M. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, *Austin*, *USA*, *1–4 November 2016*; Association for Computational Linguistics: Austin, 2016; pp. 551–561. doi: 10.18653/v1/D16-1053
183. Xu, K.; Ba, J.L.; Kiros, R.; et al. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, *Lille*, *France*, *6–11 July 2015*; JMLR.org: Lille, 2015; pp. 2048–2057.
184. Ramachandran, P.; Parmar, N.; Vaswani, A.; et al. Stand-alone self-attention in vision models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, *Vancouver*, *Canada*, *8–14 December 2019*; Curran Associates Inc.: Vancouver, 2019; p. 7.
185. Bansal, V.; Foresti, G.L.; Martinel, N. Cloth-changing person re-identification with self-attention. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, *Waikoloa*, *USA*, *4–8 January 2022*; IEEE: Waikoloa, 2022; pp. 602–610. doi: 10.1109/WACVW54805.2022.00066

186. Luong, T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, *Lisbon*, *Portugal*, *17–21 September 2015*; Association for Computational Linguistics: Lisbon, 2015; pp. 1412–1421. doi: 10.18653/v1/D15-1166
187. Yu, S.Q.; Tan, D.L.; Tan, T.N. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, *Hong Kong*, *China*, *20–24 August 2006*; IEEE: Hong Kong, China, 2006; pp. 441–444. doi: 10.1109/ICPR.2006.67
188. Iwama, H.; Okumura, M.; Makihara, Y.; *et al*. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Trans. Inf. Forensics Secur.*, **2012**, *7*: 1511−1521.
189. Rao, H.C.; Wang, S.Q.; Hu, X.P.; *et al*. A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2022**, *44*: 6649−6666.
190. Chen, Y.; Xia, S.X.; Zhao, J.Q.; *et al*. Adversarial learning-based skeleton synthesis with spatial-channel attention for robust gait recognition. *Multimedia Tools Appl.*, **2023**, *82*: 1489−1504.
191. Ben, X.Y.; Zhang, P.; Meng, W.X.; *et al*. On the distance metric learning between cross-domain gaits. *Neurocomputing*, **2016**, *208*: 153−164.
192. Ma, H.D.; Liu, W. A progressive search paradigm for the internet of things. *IEEE MultiMedia*, **2018**, *25*: 76−86.
193. Montero-Odasso, M.; Schapira, M.; Soriano, E.R.; *et al*. Gait velocity as a single predictor of adverse events in healthy seniors aged 75 years and older. *J. Gerontol. Ser. A*, **2005**, *60*: 1304−1309.

Citation: Rahi, B.; Li, M.; Qi, M. A review of techniques on gait-based person re-identification. *International Journal of Network Dynamics and Intelligence*. 2023, 2(1): 66−92. doi: 10.53941/ijndi0201005