TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Towards automatic long term Person Re-identification System in video surveillance

Athira Muraleedharan Nambiar

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-supervisors: Doctor Jacinto Carlos Marques Peixoto do Nascimento
Doctor José Alberto Rosado dos Santos Victor

Thesis approved in public session to obtain the **PhD Degree in Electrical and Computer Engineering**

Jury final classification: **Pass with Distinction and Honour**

2017

B

TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Towards automatic long term Person Re-identification System in video surveillance

Athira Muraleedharan Nambiar

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-Supervisors: Doctor Jacinto Carlos Marques Peixoto do Nascimento
Doctor José Alberto Rosado dos Santos Victor

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering

Jury final classification: Pass with Distinction and Honour

Jury

Chairperson: Doctor Mário Alexandre Teles de Figueiredo, Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee:

Doctor François Brémond, INRIA Institut National de Recherche en Informatique et en Automatique, Sophia-Antipolis, France

Doctor Hugo Pedro Martins Carriço Proença, Faculdade de Engenharia, Universidade da Beira Interior

Doctor Ana Luísa Nobre Fred, Instituto Superior Técnico, Universidade de Lisboa

Doctor Alexandre José Malheiro Bernardino, Instituto Superior Técnico, Universidade de Lisboa

Doctor Paulo Luís Serras Lobato Correia, Instituto Superior Técnico, Universidade de Lisboa

FUNDING INSTITUTIONS:

This research has been made possible with funding from the Portuguese Government (Fundação para a Ciência e a Tecnologia, doctoral grant SFRH/BD/97258/2013, project grants [UID/EEA/50009/2013], Augmented Human Assistance AHA [CMUP-ERI/HCI/0046/2013], SPARSIS [PTDC/EEIPRO/0426/2014] and from the European Commission POETICON++ (FP7-ICT-288382).

2017

Abstract

Person Re-Identification (Re-ID) is one of the most interesting tools in the realm of intelligent video-surveillance. Re-ID consists in recognising whether an individual has already been observed and then associating his identity, over a network of cameras. Person Re-ID encounters a multitude of application scenarios in real world, e.g., (1) off-line person retrieval (all the video-sequences showing an individual of interest whose image is given as query), (2) on-line pedestrian Re-ID and tracking over multiple cameras, (3) on-line authentication system via Re-ID. Traditionally, the classical approaches in Re-ID consist in exploiting the appearance cues such as color or texture of clothing. However, they restrain the system from long term applications, since those features undergo drastic variations over long periods. Hence, a new trend in Re-ID is to leverage biometric traits called soft-biometrics, like body shape and gait. Their advantage over the appearance cues is increased stability over long periods which allows long term Re-ID applications. The general objectives of this thesis is to identify good features to be used for long term application and develop novel technology towards the long term automatic Re-ID paradigm. In particular, this thesis delved beyond state-of-the-art in four main aspects.

First, a novel anthropometry based person Re-ID is proposed, by employing shape context (SC) descriptor extracted on the head-to-torso region on frontal human silhouettes, because the upper torso region of the body presents less temporal variance and occlusions with respect to other body parts thus producing more stable features. In the same work, we propose a framework for person Re-ID, using natural human compliant labels known as soft biometric traits. The generation of a novel synthetic dataset of virtual avatars (rendered by computer graphics engines), is proposed to circumvent he need for time consuming manual labelling of human datasets. Second, a novel methodology of re-identifying people in frontal video sequences, based on a spatio-temporal representation of the gait using optic flow features, is proposed. The presented methodology was evaluated in different datasets, including the novel High Definition Analytics (HDA) dataset developed in-house, with applications to re-identification in camera networks. Third, a view-point invariant person re-identification (Re-ID) by multi-modal feature fusion of 3D soft biometric cues has been proposed by leveraging both anthropometric and gait features. Two experiments were carried out under that work: (i) an extensive study of the influence of various features in the Re-ID problem and (ii) an actual demonstration of the view-point invariant Re-ID paradigm, by analysing the subject data collected in different walking directions. Finally, a context-aware ensemble fusion framework based on soft-biometric features, for long term person re-identification (Re-ID) in wild surveillance scenarios is proposed. Since biometric feature extraction is strongly influenced by the viewpoint, we associate

context to the viewing direction, and choose the best features for each viewpoint (context).

This work addresses key issues in Re-ID, with a strong focus on automated long term applications, thus strongly contributing towards the wide applicability of re-identification systems in practical real-life scenarios.

Key words: *video surveillance, person re-identification, soft-biometrics, human gait, context analysis.*

Resumo – Abstract in Portuguese

A re-identificação de pedestres é uma das áreas mais importantes no domínio de sistemas de vídeo vigilância. A re-identificação consiste em reconhecer se um pedestre já foi observado numa rede de câmaras e posteriormente associar a sua identidade numa outra câmara da rede. A re-identificação de pedestres tem muitas aplicações em cenários reais, como por exemplo, (1) na obtenção da identificação de um pedestre numa sequência de vídeo, (2) re-identificação e seguimento em tempo real de pedestres usando múltiplas câmeras, (3) sistemas de autenticação em tempo real usando re-identificação. Tradicionalmente, as aplicações clássicas de re-identificação exploram a aparência da cor ou da textura da roupa. Contudo, estes métodos apresentam limitações em aplicações de longa duração, uma vez que as características acima indicadas têm grande variações significativas ao longo do tempo. Assim, uma nova estratégia para re-identificação baseia-se no uso de atributos de biometria chamados suaves, como por exemplo a forma do corpo e do modo de andar (marcha) do pedestre. A vantagem da biometria suave face às características de aparência é que estas são mais estáveis em períodos de tempo longos, sendo úteis às aplicações de re-identificação de longo prazo. O foco principal desta tese é identificar características de pedestres que possam ser usadas em aplicações que envolvem longos períodos de tempo e o desenvolvimento de novas tecnologias para o contexto de re-identificação de longo prazo. Em particular, esta tese contribuiu para o estado da arte em quatro aspectos principais.

Primeiro, é proposto um método para re-identificação baseado na antropometria do pedestre, usando o descritor "contexto de forma" (*shape context*). O descritor é obtido na região da cabeça e do torso da silhueta frontal do corpo do pedestre, porque a região da parte superior do tronco do corpo apresenta menos variabilidade temporal e oclusões, comparado com outras partes do corpo, por isso fornece características mais estáveis. No mesmo trabalho, propomos um método para re-identificação usando características biométricas suaves do corpo humano. Propõe-se a geração de uma nova base de dados sintética de avatares virtuais (gerados por software de jogos em computador), para evitar a necessidade de efectuar a anotaçao manual de bases de dados de humanos, o que pode ser muito dispendiosa em termos de tempo. Em segundo lugar, é proposta uma nova metodologia para re-identificação de pedestres em sequências com vista frontal, baseada numa representação espaço-temporal da marcha, obtida a partir de características de fluxo óptico. Este método foi avaliado em conjuntos de dados diferentes, incluindo a nova base de dados HDA, desenvolvida por nós para aplicações de re-identificação que integram redes de câmaras. A terceira contribuição consiste em métodos de

re-identificação invariante ao ponto de vista da aquisição, usando fusão de características 3D antropométricas e de marcha. Dois estudos experimentais são apresentados: (i) um estudo alargado, que permite verificar a influência de diferentes características na re-identificação de pedestres, (ii) uma demonstração real da invariância do ponto de vista da aquisição no processo de re-identificação, analisando os dados adquiridos em diferentes direcções no deslocamento dos pedestres. Finalmente, propõe-se um método de fusão baseado em contexto e características de biometria, para a re-identificação a longo termo em cenários de vídeo vigilância. Uma vez que a obtenção das características de biometria é fortemente influenciada pelo ponto de vista da aquisição, o contexto é associado à direcção do movimento do pedestre, procedendo-se posteriormente à escolha das melhores características para cada contexto.

Este trabalho estuda questões relevantes na re-identificação, dando ênfase em aplicações automáticas de longo termo. Deste modo, fornecem-se contribuições significativas que podem ser utilizadas em cenários reais e práticos com uma ampla aplicabilidade.

Palavras-chave: *video vigilância, re-identificação, biometria, marcha, análise de contexto.*

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Alexandre Bernardino, for his immense support and guidance throughout my PhD years. He has been a great mentor, a collaborator, a guide, and a friend. This thesis would not have been completed without his commitment and encouragement, which not only influenced in the content of the thesis, but also in various aspects of my professional and personal development. Thanks a lot for being an amazing Professor. I am also grateful to my co-supervisors Prof. Jacinto Nascimento and Prof. Jose Santos Victor for all their efforts and unconditional support, during the period of my research. I can never forget the support and helping hands lent by all my professors towards me in scheduling weekly meetings, prompt reviews on the articles with valuable comments, enabling me to rehearse my presentations etc., despite the amount of work at their end, reflecting their commitment and devotion.

This research has been made possible with funding from the Portuguese Government (Fundação para a Ciência e a Tecnologia, doctoral grant SFRH/BD/97258/2013, project grants [UID/EEA/50009/2013], Augmented Human Assistance AHA [CMUP-ERI/HCI/0046/2013], SPARSIS [PTDC/EEIPRO/0426/2014] and from the European Commission POETICON++ (FP7-ICT-288382).

I am also grateful to Prof. Enrico Magli, my Master thesis adviser, who introduced me to the interesting world of research and encouraged me to pursue my studies further. My acknowledgement would be incomplete without paying thanks to Prof. Paulo Correia and Prof. Luis Ducla Soares who paved the initial stepping stones of my research life, with a challenging yet intersting topic of interest. Also, a token of gratitude to Prof. Ana Fred for her guidance and help in our collaborative research during PhD. At this point, I also express my esteem gratitude to all the teachers of my life starting from the kindergarten, who have made great impressions on me in their own unique ways, enabling me to reach in this position where I am now.

I would like to thank all my friends and colleagues from Vislab and ISR for making it a fun and interesting place to work. Much respect to my office mates Dario, Matteo, Giovanni and Pedro for encouraging me during these past four years, by sharing many interesting discussions and work time fun. Special thanks to Pedro, Giovanni and Jacinto for helping me to prepare my translated (Portuguese) thesis abstract. I also wish to thank Rui, Hugo, Nuno Montinho, Nino, Lorenzo, Prof. Gaspar, Plinio, Nuno Monteiro, Atabak, Ricardo, Afonso, Francesca, Cristian, Mehmut, Sofia, Goncalo, Beatriz, Sabina, Joao for the lively working atmosphere. My special thanks to Ricardo Nunes for his prompt technical advices and aids. I would also like to thank Ana Santos for bearing with me for all the administrative chores. Also, my colleagues

and friends of my early workspace IT lab, are also so close to my heart. I can never forget all my good friends (its a long list) in Lisbon for making my stay so much more pleasant. All of them were like a second family to me and were always ready to help me.

I would like to thank Almighty, my whole family for their support, absolute confidence in me and the unconditional love, particularly to my parents, sister and in-laws. My final yet special deepest gratitude to my loving husband Suresh Rajendran and my daughter Heera Suresh Nambiar, who are my everything. I know how much they had to manage the life during my PhD time, by compromising with many spoiled weekends, and self-cooked dinner times. My girl made me a proud working mother, when I started my research in Vislab when she was 6 months old. *Let me kindly dedicate this work to both of them.*

I must say that the list is incomplete. I thank one and all for their blessings and support throughout my life.

List of Acronyms

AEI Active Energy Image

BB Bounding Box

BF Biometric Features

CASIA Institute of Automation, Chinese Academy of Sciences

CMC Cumulative Matching Characteristic

DET Decision Error Trade-off

FAR false acceptance rate

FDEI Frame Difference Energy Image

F Frontal

FL/FS Feature Level fusion with Feature Selection

FL/NFS Feature Level fusion with No Feature Selection

FRR false rejection rate

FS Feature Selection

GEI Gait Energy Image

GEnI Gait Entropy Image

GHEI Gradient Histogram Energy Image

GHI Gait History Image

GMI Gait Moment Image

GT Ground Truth

HDA High Definition Analytics

HOF Histogram Of Flow

HOFEI Histogram Of Flow Energy Image

HOG Histogram Of Gradients

ID identifier

L Lateral

LD Left Diagonal

LL Left Lateral

MSE Mean Squared Error

NN Nearest Neighbor

PD Pedestrian Detection

PD+REID fully automated Re-Identification

RD Right Diagonal

RL Right Lateral

ROC Receiver Operating Characteristics

Re-ID Re-Identification

RMSE Root Mean Squared Error

SC shape context

SDK Software Development Kit

SL/FS Score Level fusion with Feature Selection

SL/NFS Score Level fusion with No Feature Selection

SFS Sequential Forward Selection

SVR Support Vector Regression

Contents

Chapter 1

Introduction

The journey of a thousand miles begins with one step.

— Lao Tzu

1.1 Motivation & Context

With the increase in security and forensics concerns, surveillance camera networks are unprecedentedly proliferating in both public and private areas including airports, railway stations, university campuses, shopping complexes, housing apartments, supermarkets and workplaces. In addition to providing video footage of event occurrences, surveillance cameras also act as a visible deterrent to criminals. Usually, they cover vast areas with non-overlapping fields of views. Besides, cutting edge technologies have enabled the so-called *smart security system*, which allows remote access from our smartphone, tablet or desktop, to either the home security cameras or the CCTV's at work.

As the video technology revolution brought not only cheaper access to the multimedia systems, but also large security threats on society, the number of surveillance cameras also have exceeded exponentially during the last decade. Only in the UnitedKingdom, there are between 4 million and 5.9 million CCTV surveillance, accordingto a new report from the British Security Industry Association (BSIA); one for everyeleven people [Barrett, 2013]. Each Londoner is caught on camera on average 200 timeseach day, which reveals the real influence of surveillance systems on our daily lives [Wiegler, 2008; Shitrit *et al.* , 2014; Berclaz *et al.* , 2011]. The manual analysis of the extensive collection of acquired data by the security officials is quite laborious, expensive, time-consuming and error-prone. Instead, the automated monitoring can improve the data analysis speed and the quality of surveillance [Tu *et al.* , 2007; Gala & Shah, 2014].

The analysis of data collected in surveillance camera networks serves a significant role in the evaluation and comprehension of the behaviour and activities of people. It enables to preempt suspicious events and to provide real-time alarms and situational awareness to the security personnel. The security paradigm can shift from reaction/ investigation of incidents to a more

pro-active prevention of potentially catastrophic events [Hampapur *et al.* , 2003]. Fig. 1.1 shows a typical scenario of a security officer operating on a surveillance camera network, in which he can report alarms and take immediate actions while observing intruders in the scene. This kind of proactive steps not only serve towards public safety, but also act as primary evidence in identifying the criminals, as in the 7/7 London Bombings (2005), or in Boston Marathon bombing terrorist attack (2013).

Figure 1.1: A security guard monitoring the CCTV surveillance network, and reporting the alarm in an industrial environment (Courtesy: City National Security [Sec, n.d.]- *permission requested*)

The advances in computer vision, as well as machine learning techniques in the recent years, have ameliorated this expedition towards smart surveillance at a fast pace and as a result, a plethora of algorithms for the automatic analysis of the video sequences have been proposed. They include, for instance, person detection, person tracking, activity monitoring, and person re-identification. Some survey papers such as [Gavrila, 1999b], [Gowsikhaa *et al.* , 2014] have presented them in detail.

Person detection is the process of detecting and localising each person in the images, represented via bounding boxes. This subject itself has been subject to an intensive research, see for instance [Benenson *et al.* , 2012, 2014; Dollár *et al.* , 2009; Luo *et al.* , 2014; Hosang *et al.* , 2015] that describe the state-of-the-art techniques. Regarding person tracking, the movement of the individual from one frame to another is tracked to find out the temporal consistency and the path followed in the scene (see refs [Shitrit *et al.* , 2014; Berclaz *et al.* , 2011; Niu *et al.* , 2003; Siebel & Maybank, 2002]). It deals with maintaining the accurate representation of the state and position (both the 3D space and 2D camera image plane) of the subject given measurement [Forsyth & Ponce, 2002]. Activity recognition is yet another interesting topic in video surveillance (e.g. [Hu *et al.* , 2008; Robertson & Reid, 2006a; Nascimento *et al.* , 2013]). The goal of human activity recognition is to analyse automatically ongoing activities from a video (i.e. a sequence of image frames) [Aggarwal & Ryoo, 2011]. The proposed works can be framed in either short range [Hu *et al.* , 2004a; Aggarwal & Park, 2004; Gavrila, 1999b; Moeslund. *et al.* , 2006] or far field [Robertson & Reid, 2006b] settings, depending on which the pedestrian is near or far regarding the position of the camera. Excellent reviews concerning the taxonomy of the methodologies proposed in the context are available in [Hu *et al.* , 2004b; Gavrila, 1999a; Aggarwal & Ryoo, 2011].

1.2 Re-identification

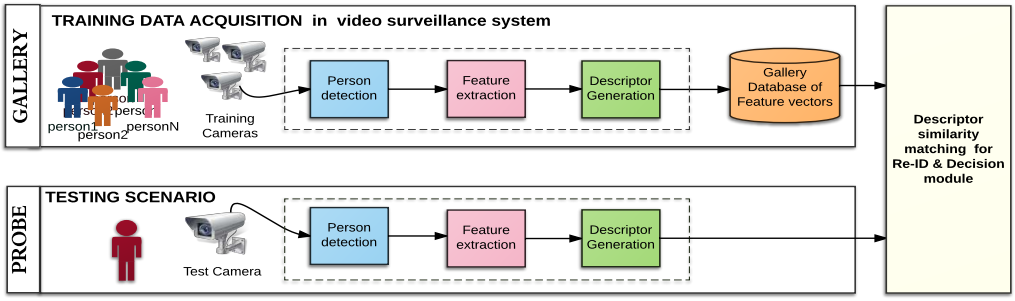


Figure 1.2: A classical person re-identification (Re-ID) diagram.

Person re-identification (Re-ID) is one of the very interesting and intricate problems. One of the earliest definitions of person re-identification owes to metaphysics Plantinga [1961], where Alvin Plantinga provided the definition to Re-ID in 1961 while discussing the relationship between mental states and behavior, as "*To re-identify a particular, then, is to identify it as (numerically) the same particular as one encountered on a previous occasion*". Afterwards, many works have been encountered in various fields such as psychology, logic, computer vision etc. Zheng *et al.* [2016]. From vision and surveillance point of view, person Re-ID is a hot topic with a high research and application significance, where the system has to *re-identify* persons in camera networks, under unconstrained conditions.

When a person disappears from the view of one camera and then reappears in another in the surveillance network, the system should be able to determine that the person has been seen before. This process of establishing connections and thus extending the tracking beyond 'blind gaps'[1]. [Doretto *et al.* , 2011] is known as Re-ID. Or in other words, re-identification is the process of establishing correspondences between images of a particular person taken from different cameras [Gala & Shah, 2014]. In Re-ID, the detected bounding box of the subject is combined with a unique label that identifies each, so that the same person at different instances will be re-identified with that unique identifier (ID).

Person Re-ID is an important task either for on-line tracking of an individual over a network of cameras or for off-line retrieval of all videos containing a person of interest. In contrast to the tracking problem, in Re-ID, the continuity constraints are much relaxed. As per mentioned in [Vezzani *et al.* , 2013], the distinction between tracking and re-identification is quite narrow and fading. Different from people tracking, the re-identification task aims to match people instances during a time delay or a change in point of view. Also, [Vezzani *et al.* , 2013] highlights that Re-ID is the best approach for associating different images of individuals, captured without a sufficient temporal or spatial continuity.

Fig. 1.2 shows a diagrammatic representation of a typical person Re-ID system. A typical pipeline of a Re-ID system includes person detection, feature extraction and descriptor matching. In the training phase, the video sequences for all individuals appearing in the surveillance

[1]Blind gaps are the gaps when the subject has left one camera and then reappeared in another camera

scenario are acquired via one or more cameras in the network. This set is denoted as "Gallery". From those image sequences, the region of interest (i.e., people) are detected. Afterwards, feature extraction is carried out and robust feature descriptors are generated. Features are the values derived from the original data, intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps. When the input data to an algorithm is too large to be processed and is suspected to be redundant, then it can be transformed into a reduced set of features named a features vector/ feature descriptor. These extracted feature vectors are stored in a gallery database for later use. Whenever a test person (probe) enters into the system in a different camera or at a different time, his feature vector is generated in the same way explained for gallery feature vectors. Then, the probe feature descriptor will be compared against the gallery database of feature descriptors by some similarity matching/ classification technique. At this stage, the Re-ID decision is made and the re-identified person ID is retrieved.

1.2.1 Re-identification: Identification vs Recognition

Re-ID is classically applied to two problems: identification and recognition. The European Commission EUROSUR-2011 [Fro, 2011] defines identification as the process to establish the unique identity of the object (name, number), as a rule without prior knowledge. The definition for recognition is to establish that a detected object is a specific pre-defined unique object [Fro, 2011]. In line with those definitions, re-identification is found to be lying in between identification and recognition [Vezzani *et al.* , 2013].

In the identification task, the goal of Re-ID is to match different observations of people using an unsupervised strategy without prior knowledge. The applications of Re-ID towards identification are witnessed in situations like long-term trajectories in wide area surveillance, open-set identification, where Re-ID helps in avoiding the identity switching, erroneous split and merge of tracks, over and under-segmentation of traces [Vezzani *et al.* , 2013]. Similarly, Re-ID is also associated with the recognition task whenever a particular query with a target person is provided, and its corresponding instances are searched in an extensive database. The result of such a query would be a set of ranked items, with the hypothesis that one and only one element of the gallery will correspond to the query [Vezzani *et al.* , 2013]. In contrast to identification scenario, recognition demands the probe to be within the database (closed-set identification[2]).

However, there are certain differences between the traditional Re-ID and recognition problems, which lies in the conditions associated with them. In recognition, usually the operator has the control over most of the conditions such as camera viewpoint (often single camera), background, subject pose, illumination, the number of persons in the acquisition, chance of occlusion, to quote a few. On the contrary, in re-identification, most of the conditions are uncontrolled, e.g., changes in background and illumination over a large number of different cameras, no control on the number of people and possible occlusions, also subjects' direction

[2]Closed set identification is where every input image has a corresponding match in the database.

varies a lot.

1.2.2 Challenges of Re-ID

There are many problems while automating Re-ID in real world surveillance systems. One of the greatest is the variation in pose and appearance of the person in various cameras and in time. This variability happens due to the different camera orientation and changes in subject pose, camera resolution or visual appearance of the person itself. In addition to that, inter-camera variations in lighting conditions, the changes in the scene illumination, different camera parameters, occlusion of the body parts, impart more constraints into the automation process. Occlusions could be either caused by other people or objects of the scene, or self-occlusions caused by own body parts. This difficulty has also been addressed in the literature (see for instance [Taiana *et al.* , 2014]). Long term Re-ID is yet another challenge because, the longer the time and space separation between views, the greater the chance that people may appear with some change of clothes or carried objects in different camera views [Gong *et al.* , 2014]. To tackle these issues, we need to have a robust feature descriptor, which is scale, pose and illumination invariant. Since generally in Re-ID, data is acquired in uncontrolled environments, i.e., without the user collaboration and varying backgrounds, the extraction of coherent discriminative feature vector demands great effort. Other complementary aspects of the Re-ID also require the collection of contextual cues such as spatio temporal topology and the situation context as well. Many works in this field are only a decade old, and all of the challenges above makes Re-ID still an open problem in computer vision.

1.2.3 Typical approaches

Most of the traditional Re-ID approaches are based on the overall human appearance in the multimedia content, viz. *Apprearance based Re-ID*. They leverage visual features based on the appearance of people, determined by their clothing (color and texture) and objects carried or associated with them. The visual descriptors include either color/ texture features or local features such as key points and edges. Rich and vast literature have been conducted on these approaches in [Bak *et al.* , 2010; Doretto *et al.* , 2011; Riccio *et al.* , 2014; Bialkowski *et al.* , 2012; Liu *et al.* , 2012]. A common problem in those techniques using color for recognition is color–constancy problem. Color constancy is the ability to assign the same color to the same object under different lighting condition [Maloney & Wandell, 1986]. As a result, the color histograms representing the very same subject look different if they are captured under different lighting conditions. Another limitation of appearance-based techniques is their short term time span, during which, the appearance described by the clothing and other attributes are considered to be constant. However, if Re-ID is to be performed for many days/ weeks, the techniques above will be quite ineffective since the holistic appearance will undergo drastic variations. For such long-term scenarios, methods based on biometric traits are found to be more suitable to be applied.

1.3 Soft biometrics

Biometrics is defined as "the science of establishing the identity of an individual, based on his/her inherent physical and behavioural traits" Ross & Jain [2007]. The term *biometrics* is coined from two Greek words: *bios* meaning 'life', and *metrics* meaning 'to measure'. A biometric-based surveillance system identifies or validates the person by extracting the characteristic biometric features of the people and comparing them with the registered gallery samples (refer Fig. 1.2). Fig. 1.3 shows various biometrics commonly used in applications. The most acclaimed and popular biometrics, also known as hard biometrics, such as fingerprint, iris, face, palm print, and voice, used in access control systems, demand the necessity for well-controlled environments and detailed computational processing, which is difficult to attain in a real-world surveillance condition. Instead, in typical video surveillance scenario, people move freely in ways that may prevent the acquisition of hard biometrics. Another genre of biometrics viz., soft biometrics sounds more promising in these scenarios.

Soft biometrics are the physical, behavioural or adhered human characteristics, classifiable in pre–defined human compliant categories which are established and time proven by humans with the aim of differentiating individuals [Dantcheva *et al.* , 2010]. Different from hard biometrics, they lack the distinctiveness and time invariance to identify a person with high reliability. However, they have certain advantages over hard biometrics, making them best suited to deploy in surveillance applications, e.g., non-obtrusiveness, acquisition from distance, non-requirement for the cooperation of the subject, computational and time efficiency and human interpretability [Nambiar *et al.* , 2016b; Reid *et al.* , 2014]. Soft biometric features leverage characteristic human traits such as anthropometric measurements, height, body size, and gait, which are more coherent for long term applications [Nixon *et al.* , 2015], than the commonly used temporary appearance cues such as dress color and texture information [Chunxiao *et al.* , 2012]. Recently, the arrival of sophisticated systems such as motion capturing devices, 3D sensors (Kinect) and high definition cameras accelerated the exploitation of soft biometrics in wide range. As a result, unprecedented many real time applications are being reported in person Re-ID and other video surveillance applications. [Gianaria *et al.* , 2014; Nambiar *et al.* , 2014a].

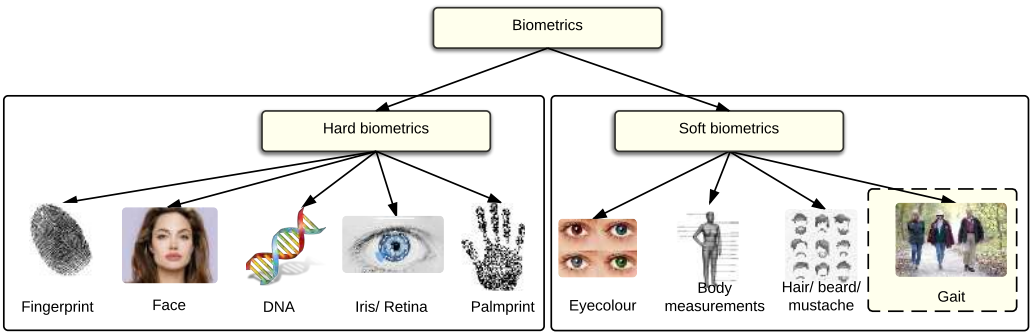


Figure 1.3: Overview of biometrics classified according to their physiological characteristics (hard biometrics) and physical, behavioural or adhered characteristics (soft biometrics). Human gait is highlighted as an instance of soft biometric.

1.3.1 Anthropometry

Anthropometry involves the systematic measurement of the physical properties of the human body, primarily dimensional descriptors of body size and shape. The first use of anthropometrics (the measurement of the human body) as a form of identification was introduced in 1883 by Alphonse Bertillon to identify recurrent criminal offenders. He first proposed a personal identification system based on biometric, morphological and anthropometric determinations [Rhodes, 1956], which known as the *Bertillonage system*. It kept records indexed by ten physical measurements: height, stretch (left shoulder to middle finger of raised right arm), bust (torso from head to seat when seated), head length (crown to forehead) and width (temple to temple), width of cheeks and the length of the right ear, left foot, middle finger and cubit (elbow to tip of middle finger). The process for obtaining each measurement was detailed within Bertillon's manual [Bertillon, 1896] and a sample of the various procedures can be seen in figure 1.4. Additional descriptions were also recorded including color of eye, hair, beard and skin, facial feature shapes, clothing, race, voice, language and any marks etc. Thus, bertillonage system is considered as one of the first sceintific recordings of anthropometric measurements.

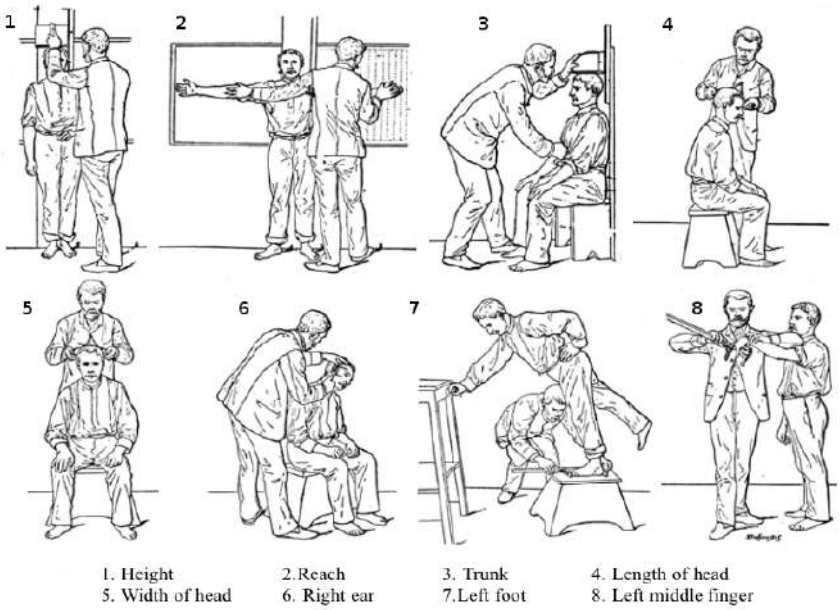


Figure 1.4: Techniques for obtaining accurate bodily measurements: Frontispiece from Bertillon's identification anthropométrique (1893), demonstrating the measurements needed for his anthropometric identification system.

This was known as the 'spoken portrait' and was recorded using a standardized shorthand. The measurements, descriptions and a standardized photograph of the individual (now known as a 'mug shot') was recorded on an identity card/ anthropometric data sheet. An example can be seen in figure 1.5. The cards were indexed in drawers each representing a specific range of the 10 metrics. This allowed hundreds of records to be quickly searched based on a set of measurements.

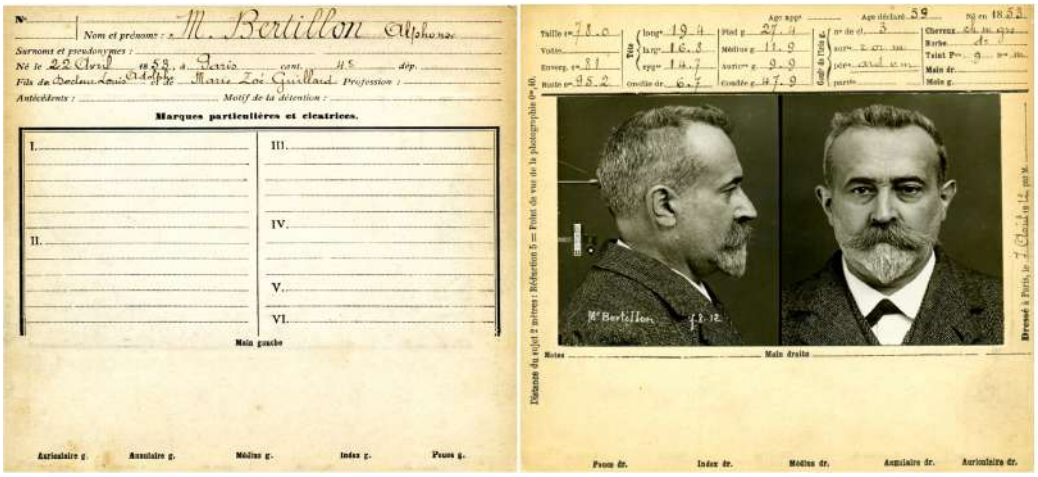


Figure 1.5: Anthropometric data sheet (both sides) of Alphonse Bertillon (1853-1914), a pioneer of the scientific police, inventor of anthropometry, first head of the forensic identification service of the prefecture de police in Paris (1893).

Today, anthropometry plays an important role in various arenas of life including industrial design, clothing design, ergonomics and architecture where statistical data about the distribution of body dimensions in the population are used to optimize products [3]. It is also employed in biometric and forensic applications as well. In terms of biometrics, it is used in computer science as a form of identification and access control as well as identifying individuals in groups that are under surveillance. A very interesting criteria of such anthropometric biometric identifiers is its human interpretability. Since soft biometric traits use human understandable descriptions (for example height, gender, hairstyle, body size), soft biometric systems can bridge the semantic gap between biometric traits and human descriptions. This presents incredible possibilities such as searching surveillance footage and databases based solely on an eyewitness' description, even without any image of the suspect.

1.3.2 Human gait

From the multitude of personal traits that characterize an individual, one of the most interesting for re-identification is human gait. It includes both the body posture and dynamics while walking Lee & Grimson [2002]. Human gait has been mentioned in many famous early works *i.e.*, Aristotle (384-322 BC) in his book "*De Motu Animalium*" on the movement of animals, and Leonardo Da Vinci (1452-1519) in his anatomic paintings. In cognitive science, gait is considered as one of the cues that humans exploit to recognize people Stevenage *et al.* [1999].

Gait is the most prevalent human movement in typical surveillance spaces. It is unique for each human and hard to fake. Several studies in neuroscience and psychology also highlight the importance of gait in human perception of the identity of others. For instance in medical situations like Prosopagnosia (face blindness), the victims use secondary cues such as gait and body appearances for person identification[4] Kress & Daum [2003]. Besides, observation of gait is believed to be an important aspect of diagnosis for several musculo-skeletal and neurological

[3] https://en.wikipedia.org/wiki/Anthropometry
[4] https://en.wikipedia.org/wiki/Prosopagnosia

conditions, such as cerebral palsy, multiple sclerosis, parkinsonism and stroke Whittle [1996].

Gait is defined as a coordinated, cyclic combination of movements that results in human locomotion [Boyd & Little, 2005]. A pictorial representation of various phases of gait is demonstrated in Fig. 1.6. The way people walk is a strong correlate of their identity. Several studies have shown that both humans and machines can recognize individuals just by their gait, given that proper measurements of the observed motion patterns are available. In cognitive science, gait is considered as one of the cues that humans exploit to recognize people [Stevenage *et al.* , 1999]. To give some background and provide the motivation for identification by gait, we refer to some early experiments in this field which came from psychophysical studies. One of the pioneering works on the peculiar nature of human motion was proposed four decades ago by Gunnar Johansson, a Swedish psychophysicist. In his famous study of biological motion Johansson [1973], using Moving Light Displays (MLDs), they instrumented the main joints of a human with bright light spots. Then, just from the observation of the motion patterns of 10-12 points, subjects reported a vivid impression of human locomotion. That work postulated that observers were able to recognize human activity (walking, running etc.) using MLDs in less than one-tenth of a second, and were able to make judgements on the gender and identity checking whether the gait pattern is familiar. Later on, follow-up studies were conducted in the paradigm by altering data acquisition conditions such as blurring the dots and relocating the position of dots Blake & Shiffrar [2007], which further confirmed that, even under indistinct conditions, motion perception is remarkably robust. In one of the famous studies Sumi [1984], a hallmark attribute associated with human motion perception was proposed that it is vulnerable to inversion. In that study, it was observed that with inverted (upside-down) MLD patterns, subjects perceived motion as very strange, despite biological.

All these studies strongly suggest that motion signals constitute valuable information from which the human brain can reliably perform detection and identification of persons, supporting the discriminative and unique nature of human gait. This has led to a large body of work being developed in the past few years towards recognition and identification of humans using gait Nixon *et al.* [2010]; Makihara *et al.* [2015]. This also accentuates the significance of gait pattern as a potential biometric tool in the surveillance application realms.

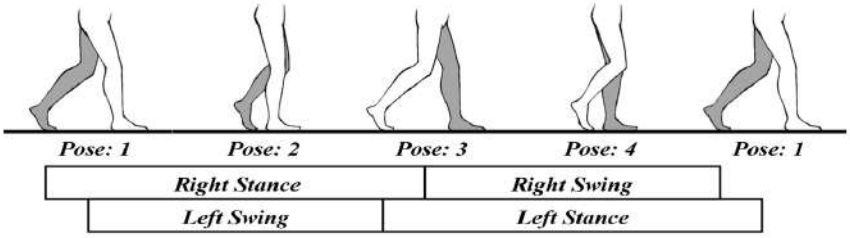


Figure 1.6: Illustrating the phases of gait. Stance Phase is the phase during which the foot remains in contact with the ground, and the Swing Phase is the phase during which the foot is not in contact with the ground. The stance phase occupies 60% of the gait cycle while the swing phase occupies only 40% of it [Loudon, 2008].

Despite the past work on gait analysis, the application of gait to re-identification only

spawned about ∼7 years ago. There are fundamental differences between the gait based recognition and Re-ID problems, which lie in the structure of the domains of application. In recognition, usually the operator has the control over most of the acquisition conditions, such as camera viewpoint (often single camera), background, subject pose, illumination, the number of persons in the acquisition, chance of occlusion, to mention a few. On the contrary, in Re-ID, most of the conditions are uncontrolled, *e.g.*, changes in background and illumination over a large number of different cameras, no control on the number of people and possible occlusions, also subjects' direction vary a lot. Furthermore, the recognition process can be generally done off-line, without any significant time constraints, whereas Re-ID often requires online response. Thus, techniques that may lead to high computational costs can not be applied. Re-ID can be viewed as a new paradigm and traditional gait recognition techniques need to be reformulated appropriately to handle it properly. Hence, due to the more realistic and unconstrained application scenarios, gait based person Re-ID has been receiving enormous attention from the computer vision and biometric communities Lee *et al.* [2014] and several works endorsed quite promising results.

For surveillance applications, gait is attractive because, it does not require active collaboration from users and is hard to fake. In addition to that, gait is unobtrusive as well as perceivable from a distance. Key advantages of gait based applications are presented in Table 1.1. Because of all these potential advantages, visual analysis of human gait for automated person recognition/ re-identification has been receiving unprecedented attention from the computer vision and biometric communities over the past few years [Lee *et al.* , 2014]. The arrival of sophisticated systems such as motion capturing devices [Josinski *et al.* , 2014], Kinect sensors [Gianaria *et al.* , 2014], high definition cameras [Nambiar *et al.* , 2014b], as well as novel machine learning techniques also have further catalysed novel researches in the field.

Nevertheless, vision-based automatic gait recognition is not without its own problems. The key advantages and challenges of the use of gait in applications are presented in Table 1.1. Gait is sensitive to certain clothing and other challenges like illness, aging, occlusions, carrying goods. In these situations, an individual can be easily distinguished for abnormality and can be better processed by a manual authentication system [Phillips *et al.* , 2002]. Also, the acquisition of good quality measures of a person's motion patterns in surveillance systems has also proved very challenging in practice. Existing technology (video cameras) suffer from changes in viewpoint, daylight, clothing and wear accessories, as well as other variations in the person's appearance. Novel 3D sensors are bringing new promises to the field, but still many research issues are open.

Gait has found forensic applications as well, for instance, securing the conviction of a bank robbery case and a burglary case based on the gait of the suspects [Bouchrika *et al.* , 2011]. Furthermore, other applications of gait analysis were also reported in physiotherapy rehabilitation [Eastlack *et al.* , 1991], medical applications [Thompson & Nutt, 2012], assessing frailty syndrome by checking the fall risk in elderly people [Kressig & Beauchet, 2006].

Being a soft biometric, gait enables the process of identity establishment even for long term scenarios, which was impossible with the traditional appearance-based Re-ID techniques. A

Table 1.1: Pros and cons of gait as a soft biometric

ADVANTAGES	DISADVANTAGES
• unobtrusive • cooperation of the user not necessary • measured at far distance • unique for each individual • cannot be easily concealed • hard to fake	• varying with illness, aging and emotional states • varying with walking surface, shoe, cloth types, carrying objects and clutter in the scenario

highlighting merit of gait in contrast to the other soft-biometrics like anthropometric or face is that gait features can encode not only the static cues but also the dynamic cues related to the movement, which is quite interesting in terms of the video surveillance because, the video sequences contain more spatial and temporal cues of the moving person than an independent images. Such dynamics as well as the spatio temporal information implicitly embedded in the multiple frames of video sequences could be exploited via the complementary gait features. The high recognition (99.84%) rate obtained using ground truth motion capturing systems [Josinski *et al.* , 2014], accentuates the potential of gait to be employed towards re-identification tasks. Different from other biometric cues, gait contains much contextual information (e.g., frequency and phase of walking, behavioural and social traits) that might be valuable to address the identity management problem.

1.4 Objectives

The research in this thesis aims to improve the performance of Re-ID system by producing an original contribution in the following areas: (a) Long term Re-ID; (b) Pose invariant Re-ID; and (c) Contextual fusion. These scenarios and research areas will be described as follows.

- 1) **Long term Re-ID :**

  The classical methods of using appearance based Re-ID schemes (i.e., color or texture) encounter the issue of short-term, where the appearance based features undergo radical changes with the change in appearance. Biometric based systems overcome such issues by exploiting robust and stable biometric descriptors, and thus motivate towards long term applications. One of the main objectives of this thesis is to analyse the influence of various soft-biometric systems such as anthropometric features and human gait, in long term [5] Re-ID. Regarding this concept, we contemplate to carry out substantial research on the discriminative biometric features of interest, various methodologies and technologies, as well as their verification in real world scenarios.

  Another interesting study we also intend to carry out in this direction is semantic bridging, i.e., to bridge the semantic gap between biometric space and use human descriptions to search a biometric database. Since human interpretability is one of the interesting criteria of soft-biometrics, human can describe our peers via verbal descriptions. For instance, an eyewitness description in a crime scene can help the authority to identify the

[5] In the context of this thesis, by *Long term*, we mean the time spanning over different days or weeks or even some months. The long term span over years could be affected by many factors such as aging, illness, dressing styles or other abnormalities. Hence, such a very long term is not considered in our work.

criminal. Semantics based Re-ID is a relatively untouched area of research and is highly desirable to be part of our research.

- 2) **Pose invariant Re-ID :**

  One of the major challenges of Re-ID is the pose of the person. Based on the direction of walking, the selected features can undergo drastic changes. Hence, in this thesis, we also would like to conduct research on pose invariant Re-ID approaches and to propose some novel ideas in this direction. This research aims to investigate ways to incorporate multi-view data to solve the problem of view-dependency.

  The state-of-the art methods leverage either some view transformation methods, invariant features or 3D methods. Among them, we intend to progress towards 3D pose invariant approaches, in which we can have view independent 3D information of the person. We aim at implementing 3D models by dint of latest emerging technologies like KINECT, Motion Capture (MoCap) technology, and then projecting the dynamics in 2D space for improving the Re-ID.

- 3) **Contextual fusion :**

  Another vital objective of this thesis work is to employ the idea of contextual fusion, i.e., to exploit the context specific information in Re-ID system. In order to facilitate this idea, we propose to enrich the system by incorporating with some contextual cues – geographical and chronological information of the video sequence, position and the orientation of the pedestrian related to the camera point (which can prioritize the appropriate modality/feature for the situation) – as well as topology of the network. These cues can help to prioritize the features based on the context. For example, when a person is far away in the field of view, gait and appearance based attributes dominates and when he is closer, part-based strategies (limbs, face) overwhelms the former. We anticipate to achieve this goal by integrating the information on spatial and temporal relationships between camera-subject or intercamera, as well as other contextual cues.

  Also, we envisage to amalgamate different biometric modalities/features and thus to make the system more robust. The rationale became evident from the prior literature review that multi-modal fusion raises the performance level of re-identification.

1.5 Original Contributions

As per the objectives mentioned in previous Section 1.4, some proposals have been already addressed during this thesis. In this section, we briefly highlight the works and scientific contributions till date. And the remaining research directions are pointed out in **Chapter7**.

During this thesis, we have already proposed some novel ideas towards soft-biometry based long-term person re-identification by dint of anthropometrics and gait. Regarding the anthropometric soft biometrics, we leverage human shape as a potential biometric cue to discriminate

people. We consider the human shape of the upper torso region, as it presents less temporal variance with respect to arms and legs motions and thus producing more stable features. In addition to that, since person re-identification is carried out in an uncontrolled environment, there are less chances for clutter and other interacting objects to make the upper body part occluded, compared to the lower parts. Similarly, another soft-biometric cue we employ in this thesis work is dynamic features. In addition to the static body features, gait features can also impart the dynamic information of the moving subject in the video sequences. Both of them have tremendous potential for immediate real-world use without upgrading the vast surveillance infrastructure. Also, two vital concepts related to Re-ID are discussed in this dissertation viz., pose-invariance and contextual cues. Many often, the information related to the direction of the data acquisition is quite significant as far as the biometric feature extraction is considered. Hence, the impact of various view-points in those feature computation and a new concept of Context-aware Re-ID leveraging the knowledge of view-points are studied in this thesis.

Original contributions resulting from the research presented in this thesis are detailed in the following section.

1.5.1 Shape based Re-ID

In this thesis, we introduce a novel descriptor for the analysis of pedestrians and its applications to person re-identification and database retrieval. A SC descriptor of the head-torso region of persons' silhouettes is shown to have a very good discrimination ability and application to Re-ID. For database retrieval using human queries, we train a map from the SC to interpretable soft biometric quantities that can be reasoned about by humans. In order to provide the best model for the regression analysis, we conducted an extensive study on the impact of various regression schemes (both linear and nonlinear) as well as cross validation schemes on shape context- biometrics pairs of our simulated dataset of virtual reality avatars. We show that a nearly linear correlation exists between SC descriptors and soft biometrics quantities in the upper human torso and illustrate its application to retrieval in databases from human queries. shape context to biometrics maps are learned from virtual avatars rendered by computer graphics engines, to circumvent the need for time-consuming manual labelling of datasets. We obtained promising results of SC based person re-identification and database retrieval from human compliant description of biometric traits, in both synthetic data and real imagery.

- *Athira Nambiar, Alexandre Bernardino and Jacinto Nascimento, "Shape context for soft biometrics in person re-identification and database retrieval", Pattern Recognition Letters, 2015.*

- *Athira Nambiar, Alexandre Bernardino and Jacinto Nascimento, "Person Re-identification based on Human query on Soft Biometrics using SVM Regression", VISAPP -11th International Conference on Computer vision Theory and Applications, Rome, Italy, 2016.*

1.5.2 Gait based Re-ID

Also in this thesis work, we analysed the impact of gait features in re-identifying people. As an initial research, we conducted a substantial overview of different approaches in gait based re-identification conducted in the past. As a byproduct, we also made a survey paper on the same topic, compiling the literature survey findings. The survey paper presents a review of the work done in gait analysis for Re-ID in the last decade, looking at the main approaches, challenges and evaluation methodologies. We identify several relevant dimensions of the problem and provide a taxonomic analysis of the current state-of-the-art. Finally, we discuss the levels of performance achievable with the current technology and give a perspective of the most challenging and promising directions of research for the future.

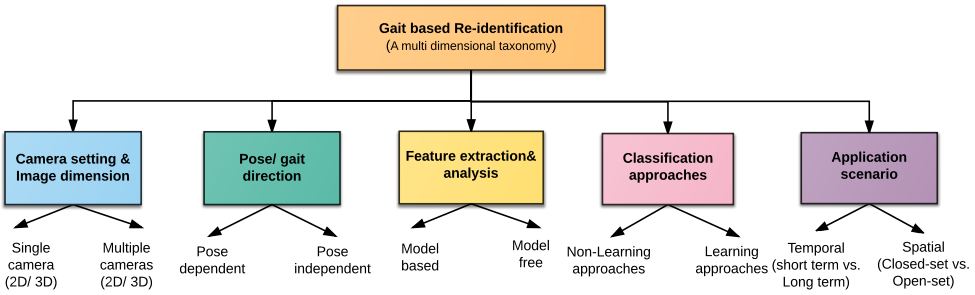


Figure 1.7: A multi dimensional overview/ taxonomic analysis of the gait based Re-identification as per carried out in our survey paper.

- *Athira Nambiar, Alexandre Bernardino and Jacinto C. Nascimento, "Gait based Person Re-identification: a Survey", acm Computing Surveys, 2017 (submitted).*

As a first proposal in gait analysis, we presented a novel methodology of re-identifying people in frontal video sequences, based on a spatio-temporal representation of the gait based on optic flow features, which we call Histogram Of Flow Energy Image (HOFEI). Optic Flow based methods do not require the silhouette computation thus avoiding image segmentation issues and enabling online Re-ID tasks. Not many works addressed Re-ID with optic flow features in frontal gait. Here, we conduct an extensive study on Institute of Automation, Chinese Academy of Sciences (CASIA) dataset, as well as its application in a realistic surveillance scenario- HDA Person dataset. Results show, for the first time, the feasibility of gait re-identification in frontal sequences, without the need for image segmentation.

- *Athira Nambiar, Jacinto C. Nascimento Alexandre Bernardino, and Jose Santos-Victor, "Person Re-identification in frontal gait sequences via Histogram of Optic flow Energy Image", Advanced Concepts for Intelligent Vision Systems ACIVS, 2016.*

1.5.3 Towards view-point invariant Person Re-ID

After studying the performance of human shape and gait in person Re-ID, we conducted an extensive study to analyse the influence of various features both individually and jointly. In

particular, we applied multi-modal feature fusion of 3D soft biometric cues viz., anthropometrics and gait. We exploited the MS Kinect^{TM} sensor v.2, to collect the skeleton points from the walking subjects and leverage both the anthropometric features and the gait features associated with the person.

In the same study, we worked towards an actual demonstration of the view-point invariant Re-ID paradigm, by analysing the subject data collected in different walking directions. In that analysis, we defined three different levels of pose-invariance, which we term as *pseudo, quasi* and *full* view-point invariance, that reflect the quantity of view points available both in the probe and in the gallery sets. Initial pilot studies were conducted on a new set of 20 people walking along four different directions, collected at the host laboratory. We illustrated, for the first time, gait-based person re-identification with truly view-point invariant behaviour, i.e. the walking direction of the probe sample being not represented in the gallery samples.

- *Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento and Ana Fred, "Towards view-point invariant Person Re-identification via fusion of Anthropometric and Gait Features from Kinect measurements", VISAPP, 12th International Conference on Computer vision Theory and Applications, Porto, Portugal, 2017.*

1.5.4 Context-Aware Person Re-ID

In the previous work, we presented a soft-biometric enabled long term Re-ID framework by exploiting human anthropometrics and gait features. We could observe that, the computation of these features depend strongly on the view-point. For instance, a person with a short stride gait is better perceived from a lateral view, whereas a person with a large chest is more distinct from a frontal view. Based on this rationale, we proposed a new framework by incorporating the information associated to the view-points (contexts), termed as 'Context-aware ensemble fusion Re-ID framework'. The major proposals of this work were (i) Model each view-point(context) with a specific set of features selected with Sequential Forward Selection (SFS) algorithm, to maximize Re-ID score in each context and (ii) Proposal of a 'Context-aware ensemble fusion framework' to fuse information from different context specific classifiers using the runtime estimate of the current context, given by an automatic context detector.

- *Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento and Ana Fred, "Context-Aware Person Re-identification in the Wild via fusion of Gait and Anthropometric features", 2nd International Workshop on Biometrics in the Wild (BWild), in conjunction with IEEE Conference on Automatic Face and Gesture Recognition, Washington DC, USA, 2017.*

- *Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento and Ana Fred, "Context-aware Person Re-identification via Fusion of Anthropometric and Gait Features", One day BMVA Technical Meetings- Security and Surveillance, British Computer Society, London, UK, 2017.*

- *Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento, Ana Fred and Jose Santos-Victor,"A context-aware method towards view-point invariance in 'in-the-wild'*

long-term Re-identification", Special Issue on Biometrics in the Wild, Image and Vision Computing, 2017. (submitted).

1.5.5 HDA Person dataset

Another contribution of this thesis work was the collaboration in the creation of a fully labelled image sequence dataset for benchmarking video surveillance algorithms. The dataset was acquired from 13 indoor cameras distributed over three floors of one building, recording simultaneously for 30 minutes. The dataset was specially designed and labelled to tackle the person detection and re-identification problems. Around 80 persons participated in the data collection, most of them appearing in more than one camera. The dataset is heterogeneous: there are three distinct types of cameras (standard, high and very high resolution), different view types (corridors, doors, open spaces) and different frame rates. This diversity is essential for a proper assessment of the robustness of video analytics algorithms in different imaging conditions. We illustrate the application of pedestrian detection and re-identification algorithms to the given dataset, pointing out important criteria for benchmarking and the impact of high-resolution imagery on the performance of the algorithms. More detais on the dataset is available from the host institution's dataset web page [6].

- *Athira Nambiar, Matteo Taiana, Dario Figueira, Jacinto Nascimento and Alexandre Bernardino, "A Multi-camera video dataset for research on High-Definition surveillance", International Journal of Machine Intelligence and Sensory Signal Processing, Special Issue on Signal Processing for Visual Surveillance, Inderscience Journal, 2014.*

I acknowledge with many thanks all the participants in the recording and labelling sessions especially, Matteo Taiana and Dario Figueira. In collaboration with them, some more contributions also have been made towards fully automated person Re-ID. In those works, we proposed an architecture for fully automated person re-identification in camera networks. Most works on Re-ID operate with manually cropped images both for the gallery (training) and the probe (test) set. However, in a fully automated system, Re-ID algorithms must work in series with person detection algorithms, whose output may contain false positives, detections of partially occluded people and detections with bounding boxes misaligned to the people. These effects, when left untreated, may significantly jeopardise the performance of the re-identification system. To tackle this problem we proposed modifications to classical person detection and re-identification algorithms, which enable the full system to deal with occlusions and false positives. We show the advantages of the proposed method on a fully labelled video dataset acquired by 8 high-resolution cameras in a typical office scenario at working hours. Since these works are collaborative tasks and not the core of this thesis, are not explained in this dissertation. Nevertheless, the papers resulting from this research are listed below:

- *Matteo Taiana, Dario Figueira, Athira Nambiar, Jacinto Nascimento and Alexandre Bernardino, "Towards Fully Automated Person Re-Identification ", VISAPP 2014, 9th*

[6] http://vislab.isr.ist.utl.pt/hda-dataset/

International Conference on Computer vision Theory and Applications, Lisbon, Portugal, January, 2014.

- *Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento and Alexandre Bernardino, "The HDA+ dataset for research on fully automated re-identification systems", Proc. of ECCV2014 Workshop on Visual Surveillance and Re-identification, Zurich, Switzerland, 2014.*

1.6 Outline of thesis

The remaining chapters of this thesis are structured as follows:

- **Chapter2** provides a detailed literature review of existing shape and gait based re-identification tasks. Under the gait analysis literature, detailed survey of techniques used for feature extraction as well as robust pose invariance are also described. In addition to that, various datasets and the evaluation metrics used for Re-ID, are also detailed.
- **Chapter3** describes the implementation of the shape based person Re-ID framework used in this thesis, with a shape context based regression approach. The system architecture, methodology and the experimental results for both nonlinear and linear regression analysis schemes are explained in this chapter.
- **Chapter4** illustrates our work of gait based Re-ID leveraging optic flow features, in the frontal views. As a part of this, generation of a novel feature vector called Histogram Of Flow Energy Image (HOFEI) is described by fusing the Histogram Of Flow (HOF) into the Gait Energy Image (GEI) baseline architecture. System framework, methodology used and the experimental results on two datasets (CASIA & HDA) are presented here.
- **Chapter5** deals with the multi-modal fusion as well as pose-invariant Re-ID studies carried out in this dissertation. Regarding the former, a thorough study of the impact of biometric features (i.e., anthropometric and gait features) in Re-ID is carried out both individually and jointly. To deal with the latter, a benchmark assessment is conducted by experimenting with different view-points in the probe and gallery samples.
- **Chapter6**, provides another interesting aspect of context-aware re-identification, where we incorporate the information associated to the view-points (contexts) and thus proposes a novel 'Context-aware ensemble fusion Re-ID framework'. In the studies conducted in **Chapter5** and **Chapter6**, we exploited MS Kinect^{TM} v.2 based indoor person Re-ID set up (a new pose invariant dataset collected in house) as the test bed,by leveraging 3D skeleton joints of the subjects.Several case-studies were conducted and the experimental results are explained in detail in the respective chapters.
- **Chapter7** concludes this dissertation with a summary of the research and a discussion of the advantages and limitations of the work. It also provides some suggested directions for future research in this area.

- **Appendix** demonstrates our HDA dataset generation (for video surveillance benchmarking applications) and KS20 Vislab Multi-view Kinect Skeleton dataset acquisition procedures.

Chapter 2

Literature Review

If I have seen further, it is by standing on the shoulders of Giants.

— Sir Isaac Newton

Pioneering research in video surveillance applications using soft biometry has been witnessed in the last decade, which was enhanced with the introduction of advanced motion capturing devices and high definition cameras. In this chapter, we mainly discuss two soft biometrics - anthropometry and gait - and their major state-of-the-art researches towards person re-identification.

2.1 Anthropometry based Re-ID: Related works

Physical body descriptions (anthropometric features) have also been used in biometric techniques as an ancillary data source where they are referred to as soft biometrics, as opposed to primary biometric sources such as iris, face etc. As mentioned earlier, soft biometric traits lack the distinctiveness and permanence to accurately identify a person, in contrast to the hard biometrics. In order to cope with this issue of any single soft biometric, fusion of multiple soft-biometric traits have received huge acclaim in the biometric and computer vision communities. The basic idea behind is that by agglomerating many soft-biometric features could construct a reasonably unique signature. For example, in [Dantcheva *et al.* , 2010], a bag of soft biometric traits (e.g., facial and body soft biometrics) was presented for person re-identification. They proposed a general framework by integrating both the primary biometrics (i.e. face, iris) and soft biometric system (i.e. height, gender)- and thus reinforce the signature uniqueness. Similarly, [Barbosa *et al.* , 2012a] presented a set of 3D soft biometric cues related to anthropometric measurements, obtained from KINECT RGB-D sensors and employed in person re-identification. The retrieval using soft biometrics is also addressed in [Reid & Nixon, 2011], where they proposed a method of comparative human descriptions for soft biometrics.

The major advantages of the soft biometrics over hard biometrics are two fold: they can be acquired even at a distance without the subject collaboration, and they are human interpretable and hence can bridge the semantic gap between biometric traits and human descriptions. In this thesis, we focus on the latter characteristic feature of soft biometrics which enables the

system to retrieve the person right away from the human verbal description of the person. [Sridharan *et al.* , 2005] proposed a facial image retrieval system that is queried using verbal descriptions. Queries can include up to 14 defined features, composed of five Boolean descriptors (e.g. presence of beard) and nine categorical labels (e.g. nose width, face length, hair colour). In another work, [Samangooei & Nixon, 2010] developed a soft biometric system which identifies subjects from video footage (Soton gait database). This description was composed of 23 absolute categorical labels of the soft biometric traits /categorical attributes, like hair colour, hair length, height, hip length, chest, arm length, etc., and they were described using absolute labels. [Reid & Nixon, 2011] extended the work further by leveraging comparative human descriptions, which used visual comparisons between subjects. However all the aforementioned works required laborious manual annotations over real world dataset, by a large number of human users. We, try to automate this process by exploiting machine learning technique and modern computer graphics technology herein. Details of the experiments carried out are further explained in **Chapter3**.

In this thesis, we consider the shape information (shape context descriptor) of the upper torso as the anthropometric traits (eg., parameters of neck, head, chest) from the edge information of the silhouettes. Hence, here we particularly mention some of the prior works carried out on similar approaches in the literature.

Silhouettes: Most of the state-of-the-art methods leverage either color information or local feature descriptors inside the human body after segmenting the silhouettes. In [Truong *et al.* , 2010], a robust classification procedure exploited the discriminative nature of sparse representation to perform people re-identification. [Aziz *et al.* , 2011a] presented a person Re-ID method based on appearance classification and silhouette part segmentation using various descriptors such as SIFT, SURF and SPIN. In this thesis, instead of appearance cues, we exclusively depend upon contour information and propose a new way of long-term person re-identification using silhouettes. To the best of our knowledge no complete work for person re-identification, leveraging solely the edge information of the silhouette is reported in the literature.

Shape Context: The application of shape context (SC) in human video surveillance systems are reported in the state-of-the-art. Some works are found in pedestrian detection by [Leibe *et al.* , 2005], highlighting that *SC* descriptor trained on real edge images exhibited high performance, particularly on difficult images and backgrounds. Some application of SC have also been employed in gait recognition [Zhang *et al.* , June 2009] where SC is used to compute the similarity between two Procrustes Mean Shape, which is a compact representation of gait sequence. A similar application of SC is found in human pose estimation [Agarwal & Triggs, 2004]. However, the literature is scarce concerning the use of SC in Re-Identification (Re-ID) applications. One exception is [Wang *et al.* , 2007] that created shape labeled images by means of shape and appearance models which was inspired from the idea of shape context. Another work [Kviatkovsky *et al.* , 2013], used SC descriptors to represent the intra distribution of colors for person re-identification. In our work, we propose shape context features computed on the contour of the silhouette of frontal images of persons. Our studies on the topic is discussed in Chapter3.

2.2 Gait based Re-ID: Related works

The traditional approaches for gait feature extraction are classified into two major categories: model free approaches and model based approaches [Wang *et al.* , 2010]. Model free approaches acquire gait parameters by performing measurements directly on 2D images, without adopting specific model of human body or motion (e.g. silhouettes, optic flow, history of movements). Thus, they are simple and faster. However, they are data- driven and hence highly prone to the occlusions, pose and scale variance, camera view angle, direction of walking, clothing/appearance change [Chen *et al.* , 2009]. On the contrary, model based approach makes use of explicit gait models, whose parameters are estimated using the underlying kinematics of human motion in a sequence of images (e.g., step dimensions, cadence, human skeleton, body dimensions, locations and orientations of body parts and joint kinematics). These methods mostly focus on gait dynamics and are more resistant to problems like changes of view and scale. However, the model based methods are often computationally expensive due to the large number of parameters that need to be fitted [BenAbdelkader *et al.* , 2004], necessity for high-quality data, and other difficulties of determining the position of joints in arms and legs.

Classical model based approaches consist in structural models (usually 2D or 3D) and motion models. Structural models define the human topology as functions of the body parameters, whereas the motion models determine the kinematics of the motion of each body part. The model based techniques either model the body or the walk of the person as it will appear in the imagery. They are more robust to a variety of factors (changes in the appearance of walking person due to clothing, carrying goods, background) and typically yield better recognition results compared to model free approaches in inter-class conditions [Sivapalan, 2014]. Some acclaimed works in model based approaches are [Bobick & Johnson, 2001; Sivapalan *et al.* , 2011; Gianaria *et al.* , 2014], to quote a few.

The model free approaches, in contrast to model based approaches, don't require intermediate 2D or 3D geometric or kinematic models. Model free methods characterize the whole motion pattern of the human body by analysing the variations in the silhouette shapes or body motion over time, regardless of the underlying structure. They circumvent the difficulties in fitting models to data and are computationally simpler compared to the model based approach. Nevertheless, they are sensitive to view angle, pose and scale. Since the model fitting process is very challenging in complex backgrounds, uncontrolled environments, and occlusions, most works on these conditions use model free approaches. Some acclaimed works in model based approaches are [Han & Bhanu, 2006; Sarkar *et al.* , 2005; Goffredo *et al.* , 2008; Nambiar *et al.* , 2012], to quote a few.

In this thesis, we consider the model-free gait analysis technique, leveraging optic flow features. Hence, here we particularly mention some of the prior works carried out on similar approaches in the literature. Nevertheless, we have already conducted a rich survey on the various gait based Recognition/ re-identification schemes in our survey paper.

Energy Image: One of the most acclaimed research in model free gait recognition viz.,

GEI by [Han & Bhanu, 2006], presented the idea of generating spatio-temporal description by averaging the normalized binary silhouette over gait cycle. Afterwards, a large number of variants of GEI's were introduced, which formed the basis of many recent model free gait recognition systems e.g., Active Energy Image (AEI) [Zhang *et al.* , 2010], Gait Entropy Image (GEnI) [Bashir *et al.* , 2010], Gradient Histogram Energy Image (GHEI) [Hofmann & Rigoll, 2012], Frame Difference Energy Image (FDEI) [Chen *et al.* , 2009] etc.

Optic flow: The idea of HOF was adopted from Histogram of Oriented Gradients (HOG), which divide the image into cells and compile a histogram of gradient directions, weighed by its magnitude for the pixels within each cell [Dalal & Triggs, 2005]. The same approach has been extended to the optic flow and the spatial derivatives of its components [Dalal *et al.* , 2006; Moreno *et al.* , 2015]. Optic flow and their histograms have also been proposed for gait analysis such as in [Bashir *et al.* , 2009] by using motion intensity and direction from optical flow field, while in [Lam *et al.* , 2011] a silhouette based gait representation has been used to generate gait flow image. In the field of optic flow based gait recognition also some energy image concepts were proposed in the recent works by [Yang *et al.* , 2014] and [Lam *et al.* , 2011]. However, both of those works were reasonably insufficient to convey the motion information of the whole human body since their optic flow measurements are on the binary silhouette edges.

Much inspired from the aforementioned literature studies, here we propose a novel spatio-temporal gait representation termed as Histogram Of Flow Energy Image (**HOFEI**), which is a dense descriptor computed over the entire body parts. Different from the aforementioned literature on Optic flow based gait recognition conducted in the lateral view, we demonstrate the potential of our proposal in the front view (in HDA and CASIA dataset), for which no similar state-of-the-art using Optic flow has been reported (explained in Chapter 4). However, there have been some works in CASIA dataset frontal sequences, leveraging the binary silhouettes for gait recognition. Chen et al [Chen *et al.* , 2009] demonstrated the performance of various gait features including GEI, Gait History Image (GHI), Gait Moment Image (GMI), FDEI. in each view angle, from frontal to rear view. We will show that our proposed method is competitive with this state-of-the-art, while using an optic flow method, which does not require silhouette segmentation.

2.3 View-invariant person Re-ID: Related works

Many of the classical Re-ID systems found in the literature were built on appearance based features [Doretto *et al.* , 2011; Riccio *et al.* , 2014], exploiting the colour/ texture of the clothing. Some works discussed view-point invariant techniques by exploiting 3D scene information, pose priors etc. [Bak *et al.* , 2014, 2015; Wu *et al.* , 2015]. However, those approaches were not pertinent towards long term Re-ID applications, where the appearances change drastically. In recent years, a new trend employing biometric information has blossomed, owing to the precise and advanced data capturing machines (e.g. HD cameras, motion capture, kinect sensor), especially in analysing the 3D body information that enables view-point invariance.

Many works have been proposed towards view-point invariant Re-ID. In [Zhao *et al.* , 2006],

[Iwashita *et al.* , 2010] multiple 2D cameras were used to reconstruct the 3D volumes and thus achieve view-point invariance. Other works use multiple 2D cameras to fit 3D models in the volumetric data e.g. 3D ellipsoids [Sivapalan *et al.* , 2011], articulated cylinders [Ariyanto & Nixon, 2011] and 3D volume shape by the intersection of projected silhouettes [Seely *et al.* , 2008]. Current state-of-the-art view-point invariant techniques are presented in [Iwashita *et al.* , 2014], [Fernández *et al.* , 2016]. In [Iwashita *et al.* , 2014], a method using a 4D gait database was proposed. At each frame of a gait sequence, the observation angle is estimated from the walking direction by fitting a 2D polynomial curve to the foot points. Then, a virtual image corresponding to the estimated direction is synthesized from the 4D gait database. [Fernández *et al.* , 2016] presents a multi-view-point gait recognition technique based on a rotation invariant gait descriptor derived from the 3D angular analysis of the movement of the subject. In addition to them, many multi-view gait recognition methodologies have also been developed in the last decade, which also could be extended towards person identification application e.g., a multi-view gait recognition method using activity-specific biometrics [Johnson & Bobick, 2001], incorporating view transformation models (VTM) to facilitate the mapping among different view angles [Kusakunniran *et al.* , 2009], fusion of different feature subspaces of aperiodic feature representations [Padole & Proença, 2017], cross capture modality named BackfilledGEI(BGEI) [Sivapalan *et al.* , 2012] etc. to name a few.

Some works exploiting view-point invariant RGBD sensors (e.g. kinect) have also been proposed in the literature. In the work by [Barbosa *et al.* , 2012b], they leveraged the soft-biometric cues of a body for person Re-ID. However they used only the static body information i.e. skeleton based features and surface based features, in the frontal view. Later, some works employed the gait features as well, e.g. stride and arm kinematics [Gabel *et al.* , 2012], knee angles [Aarai & Andrie, 2013], anthropometric and dynamic statistics [Gianaria *et al.* , 2014], anthropometric and angles of lower joints [Andersson & Araujo, 2015].

In our work, we build on the aforementioned state-of-the-art works by proposing some novel ways of improving the Re-ID algorithm, in terms of feature extraction, feature fusion and impact of view angles. In particular, we examine the Re-ID accuracy of various anthropometric and gait features via both individual as well as joint schemes. In addition, we explicitly conduct a view-point invariant Re-ID scenario by collecting video sequences of people walking in different directions, whereas previous related works collect data in a much controlled predefined single direction say, frontal or lateral. Our reserach work on pose-invariant Re-ID is described in Chapter5.

2.4 Context-aware person Re-ID: Related works

The arrival of Kinect^{TM} RGBD sensor gave rise to unprecedented advancements in the biometric and computer vision community, to devise many sophisticated techniques allowing view point invariance. Many Re-ID works utilizing Kinect data have been reported in the literature. By exploiting soft-biometric cues in contrast to the primarily appearance cues (colour or texture), they promote long term person Re-ID. In one of the earlier works *viz.*, [Barbosa *et al.* , 2012b], a specific signature built from a composition of several soft biometric (*e.g.*, skeleton and surface based features) cues extracted from the depth data, was computed for each subject. Then, Re-ID was accomplished by matching these signatures against the test subjects

from the gallery set. Similarly, person re-identification from soft biometric cues was also addressed in another work [Munaro *et al.* , 2014b], where skeleton descriptors (by computing several limb lengths and ratios) and shape traits (using point cloud shape) were used in order to re-identify people. In [Gianaria *et al.* , 2014] both anthropometric features (*e.g.*, height, leg length, etc) and dynamic parameter related to gait (*e.g.*, knees movement, head oscillation) were used. Also, in [Andersson & Araujo, 2015] a methodology to extract anthropometric and gait features was addressed showing the results of applying different machine learning algorithms on subject Re-ID tasks. However, in those approaches, the acquisitions were conducted in a constrained manner *i.e.*, in a particular view-point. In our work, we build upon the state-of-the-art works but in a less constrained conditions, by explicitly imposing view-point changes in the dataset and by exploiting relevant features in each of those view-points (contexts). See Chapter6 for details.

Many definitions of context were encountered in the literature, depending on its field of application. According to the dictionary, context is defined as "*the surroundings, circumstances, environment, background or settings that determine, specify, or clarify the meaning of an event or other occurrence*" [wik, n.d.]. In our work, we define context as the view-point setting, under which features are computed. The application of context has been reported in diverse fields, for instance, in customer behaviour applications [Palmisano *et al.* , 2008], where the context is viewed as the intent of a purchase (*e.g.* context of a gift). In [Ding *et al.* , 2011], an application for Re-ID of the subject from instant messaging in a web surfing navigation is proposed. The context is the special characteristics of chatting text (*e.g.* content, token, syntax and structural based features). In [Panniello *et al.* , 2016] context was used for online customer re-identification, where the intent was to investigate whether customer behavior models of the context (in which a transaction takes place), can increase client re-identification performance. The contextual information is interpreted as the time of day when or the location where digital data was created. Few works, however, addressed the concept of context within the re-identification setting as we propose in this paper. In particular, [Zhang *et al.* , 2014] proposed a Re-ID paradigm which leveraged heterogeneous contextual information together with facial features. In particular, they used clothing, activity, human attributes, gait and people co-occurrence as various contexts, and then integrated all of those context features using a generic entity resolution framework called RelDC. Some other recent Re-ID works utilized context as a strategy for refining the classical Re-ID results via re-ranking technique [Leng *et al.* , 2015; Garcia *et al.* , 2015]. In those works, in addition to the content information of the subjects, they also paid attention to the context information (k-common nearest neighbors) to fine tune the Re-ID results. From our literature review, it was comprehended that context is a new tool whose effectiveness in Re-ID applications is yet to be completely explored.

2.5 Datasets available for soft-biometrics based Re-ID

In this section, we describe the various datasets available towards long-term person Re-identification exploiting soft-biometrics. Since we focus our research upon static and dynamic body biometric measurements (i.e., anthropometric and gait) in order to identify people, we herein mention concerned datasets in both of those areas. Section 2.5.1 refers to the gait based datasets publicly available. Since gait analysis is independent of the appearance cues and hence make it

available for long term invariance, we include all the possible unconstrained gait databses. For the anthropometry based long term Re-ID (see Section 2.5.2), we ignore all the appearance based Re-ID since they are short-term oriented, and concentrate on the datasets acquired via depth sensors. These sensors provide depth and skeleton data which are less subject to daily variations of a person's appearance than RGB images.

2.5.1 Gait based Re-ID datasets:

There are a plethora of publicly available datasets for video surveillance applications such as traffic monitoring, action recognition, subject tracking, person detection, etc. Among them, only a few can be adopted for gait based Re-ID. Since the gait based Re-ID demands a spatio-temporal evaluation of the walking pattern, gait based Re-ID datasets require multiple-shot image sequences with several complete gait cycles. The databases for gait Re-ID research usually necessitate a large number of subjects and camera angles which are captured in realistic surveillance scenarios under very general conditions. In this section, we enumerate the datasets available for gait based Re-ID, *i.e.*, those containing different view angles of walking as well as multiple-shot (continuous gait frames) video data sequences, which more faithfully represent the conditions of gait based Re-ID scenario. Hence, even though there are many gait analysis and person Re-ID data sets available publicly[1], not all of them could be used towards gait based Re-ID since they contain either controlled gait sequences (*e.g.*, MoBo [Gross & Shi, 2001], Multi Biometric Tunnel [Seely *et al.* , 2008]) or single or nonsequential frame images (*e.g.*, VIPeR [Gray & Tao, 2008], GRID [Loy *et al.* , 2009], CAVIAR4REID [Cheng *et al.* , 2011]). The nature of the background, the type of cameras employed in the framework, and the viewpoint and gait direction in the image sequences have a significant influence on the analysis algorithms and the performance evaluation.

–**CASIA**: It is one of the largest and more popular databases in gait analysis and related research[2]. It contains four different datasets: Dataset A Wang2003 is an outdoor gait dataset consisting of 20 people walking in 3 directions (lateral (90 degrees), frontal (0 degrees) and 45 degrees). Dataset B is an indoor gait dataset composed of 13640 samples acquired from 124 subjects at 11 different views Yu2006. Dataset C is collected using an infrared camera from 153 subjects at four different conditions (normal, slow, fast, normal with a bag). Dataset D was collected simultaneously with a camera and a Rscan Footscan [3] on 88 subjects. In most of the gait analysis works, Dataset A and Dataset B are widely used due to their realistic mode of data acquisition Liang2003,Sivapalan2012. Specifically for Re-ID applications, Dataset B is well suited since it addresses the issue of pose, by acquiring the scene with a network of 11 cameras, each with a view angle separation of 18 degrees. In Wei2015,Nambiar2016b,Liu2015,Lu2014 it was employed in **gait based Re-ID**.

–**SOTON**: The SOTON[4] database Nixon2006 was developed at the University of Southamp-

[1] This page collects all public datasets that have been tested by person Re-ID algorithms. https://robustsystems.coe.neu.edu/sites/robustsystems.coe.neu.edu/files/systems/projectpages/reiddataset.html

[2] http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp

[3] http://www.rsscan.com/footscan/

[4] http://www.gait.ecs.soton.ac.uk/

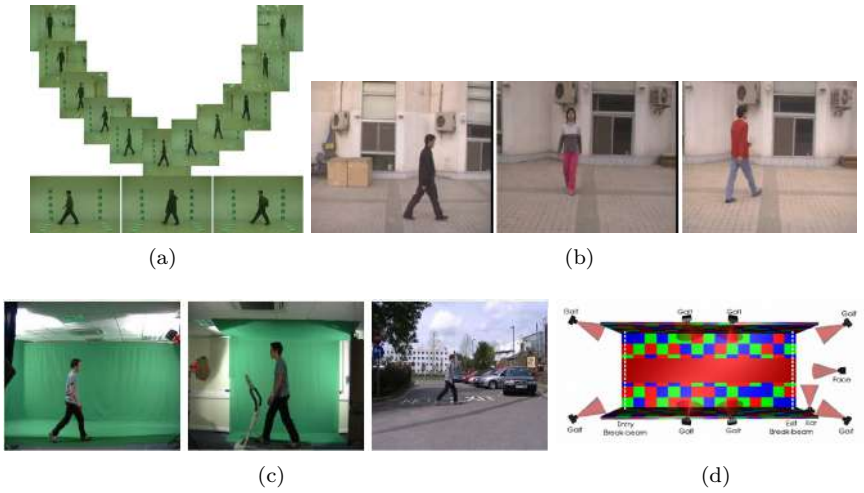


Figure 2.1: Shot examples from (a) CASIA indoor dataset (b) CASIA outdoor dataset [Wang *et al.* , 2003b] and (c) SOTON large dataset [Nixon & Carter, 2006] (d) Multi-biometric tunnel [Seely *et al.* , 2008]. Both CASIA and SOTON database contain the indoor and outdoor scenarios. Nevertheless, they are captured in highly controlled conditions, very dissimilar to a typical surveillance environment. Hence, these databases are well suited for gait recognition rather than re-identification.

ton, with the principal aim of developing new technologies for recognising people at a distance. It is composed of a large and a small database. The former consists of nearly 114 subjects and over 5000 samples but contains little variability (each subject was filmed from only two different views over three separate scenarios). The small database contains only 12 persons but is more complete regarding the covariates (change in clothing, accessories and different speeds). For the large database three scenarios are analyzed, namely outside, inside track, and inside treadmill, whereas the small database contains subjects walking with different appearances, and at various speeds, all collected in the indoor scenario. The SOTON dataset has been used for gait recognition Bashir2010,Wang2012,Wei2015 as well as **gait based Re-ID** [Wei *et al.* , 2015].

–**USF**: The USF[5] dataset Sarkar2005 contains 1870 sequences acquired from 122 subjects. It comprises elliptical movements of people walking in front of cameras. For each person, up to five covariates were manipulated such as shoe type, bag carried, type of surface, viewpoint and time instants. The data was collected over four days, in which 33 subjects were the same. Some of the works using the USF dataset are [Han & Bhanu, 2006; Zhang & Viola, 2007; Wang *et al.* , 2012]. Human identity recognition using USF database has been presented in Lu2014.

–**SAIVT**: Recently, a multi-camera surveillance database SAIVT[6] was created by Bialkowski et al. [Bialkowski *et al.* , 2012] for the evaluation of person recognition and Re-ID models in realistic surveillance scenarios. The database consists of unconstrained walking video sequences of 150 people, collected inside a building. Eight surveillance cameras acquired images

[5] http://figment.csee.usf.edu/GaitBaseline/
[6] https://wiki.qut.edu.au/display/saivt/SAIVT-SoftBio+Database

(a)

(b)

(c)

Figure 2.2: Example of video frames from (a) SAVIT [Bialkowski *et al.* , 2012], (b) AVAMVG [López-Fernández *et al.* , 2014] and (c) HDA Person dataset [Nambiar *et al.* , 2014b] . All of these datasets are collected indoor under multi-camera networks. They provide data of realistic uncontrolled conditions, with significant variation in the pose, illumination and camera view angle.

of resolution 704×576 pixels at a framerate of 25 frames per second. The dataset provides a highly unconstrained environment for testing person Re-ID models in conditions that are closer to real scenarios. **Gait based Re-ID** employing SAIVT dataset has been published in the literature [Bedagkar-Gala & Shah, 2014].

–**AVAMVG**: The AVA Multi-View Dataset for Gait Recognition AVAMVG[7] [López-Fernández *et al.* , 2014] is another recent dataset directed towards robust recognition. It collects data of 20 people walking along ten trajectories each, using six calibrated cameras with different views angles. Images have a resolution of 640 x 480 pixel and are acquired at 25Hz. The database has been specifically designed to test gait recognition algorithms based on 3D data. The binary silhouettes of each video sequence are also provided. Some works in multi-view and viewpoint-independent gait analysis using the AVAMVG dataset are [Castro *et al.* , 2014; Fernández *et al.* , 2016].

–**HDA person dataset**: HDA[8] dataset is a multi-camera video dataset mainly dedicated

[7] http://www.uco.es/investiga/grupos/ava/node/41
[8] http://vislab.isr.ist.utl.pt/hda-dataset/

to benchmarking video surveillance algorithms such as person detection and Re-ID [Nambiar *et al.* , 2014b]. It is a fully labeled image sequence dataset, collected using 13 indoor cameras for a duration of 30 minutes. More than 64,000 annotations were performed on a total of more than 75,000 frames. The dataset is quite diverse in terms of types of cameras (standard, high and very high resolution), environment types (corridors, doors, open spaces) and frame rates (5fps, 2fps, 1fps). Several of the acquired image sequences are in the HR range (1,280×800 pixel and 2,560×1,600 pixel), which makes the HDA dataset the first one to include labeled video sequences of such resolution. Extended versions of the dataset have been published *viz.*, HDA+ dataset [Figueira *et al.* , 2014] along with a novel framework towards fully automated person Re-ID [Taiana *et al.* , 2014]. Some **gait based Re-ID** works employing HDA datasets are also available in the literature [Nambiar *et al.* , 2016c; Wang *et al.* , 2016].

–**i-LIDS**: The Imagery Library for Intelligent Detection Systems i-LIDS is the U.K. government's benchmarking dataset towards video analytics systems. It comprises a library of CCTV video footage collected from various scenarios mainly categorized as event detection and object tracking scenarios. Among them, the i-LIDS multiple camera tracking (MCT) scenario was collected inside a busy hall using five cameras at 25fps. 119 people were captured, but the average image count per person is four, which is very few for gait based applications. The presence of occlusions and quite large illumination changes make this dataset very challenging for the Re-ID task. An extended versions of i-LIDS dataset, iLIDS-VID[9] is presented in [Wang *et al.* , 2014]. Re-ID research works using i-LIDS have been conducted in the recent years [Zheng *et al.* , 2009]. In particular, [Bouchrika *et al.* , 2016] presented identity tracking across multiple cameras using i-LIDS and [Wang *et al.* , 2016] presented **gait based Re-ID** using i-LIDS-VID dataset.

–**TUM-GAID**:A new freely available database or multimodal gait recognition was proposed by [Hofmann *et al.* , 2014]. It is denoted GAID[10] (Gait from Audio, Image and Depth) and contains RGB video, depth, and audio concurrently. It is composed of recordings from 305 people in three variations, making it as one of the largest to-date. A second subset of 32 people was recorded to further investigate challenges of temporal variability. Some **gait based Re-ID** research works employed TUM-GAID dataset are [Geiger *et al.* , 2014; John *et al.* , 2013].

–**PRID2011**: This dataset was created for the purpose of testing person Re-ID approaches[11] [Hirzer *et al.* , 2011]. It consists of image frames extracted from two static camera recordings, depicting people walking in different directions. Images from both cameras contain variations in viewpoint, illumination, background and camera characteristics. 475 and 856 person trajectories were recorded via individual cameras, with 245 persons appearing in both views. The dataset has two versions: a single-shot scenario and a multi-shot scenario. PRID2011 has been employed in **gait based Re-ID** applications in [Wang *et al.* , 2016].

[9] http://www.eecs.qmul.ac.uk/~xz303/downloads_qmul_iLIDS-VID_ReID_dataset.html
[10] https://www.mmk.ei.tum.de/verschiedenes/tum-gaid-database/
[11] https://lrs.icg.tugraz.at/datasets/prid/

–**PETS2009**: A widely known dataset is PETS[12] [Ferryman & Shahrokni, 2009] presented at the 2009 edition of the international workshop on performance evaluation of tracking and surveillance. It was recorded in a public space outdoor scene at University of Reading, UK. It is a multi-camera system consisting of 8 cameras, and it contains three sequences with different crowd activities in a real world environment. The partial dataset that addresses person tracking consists of three subclasses based on their subjective difficulty level, associated with the density of the crowd. Refer to [Baltieri *et al.* , 2011b] for some benchmarking results. Since the dataset provides multi-shot sequences with multiple viewpoints, this is useful towards **gait based Re-ID**. In [Bouchrika *et al.* , 2016], gait based Re-ID and tracking across multiple non-intersecting cameras have been applied to the PETS2009 dataset.

–**ETZH**: The ETZH[13] dataset [Ess. *et al.* , 2007] is a dataset originally proposed for human detection. It contains four video sequences captured with a moving stereo rig, thus presenting an additional challenge given by the moving cameras. Albeit people's pose, appearance and scene illumination have a reasonable degree of variation, most of the viewing angles are quite similar (frontal). It consists of full frames and bounding box annotations. Some works using the ETHZ dataset in Re-ID application are [Farenzena *et al.* , 2010; Zhao *et al.* , 2013]. Since full frames with multi-shot sequences are available, this dataset could be employed for **gait based Re-ID** applications.

–**3DPeS**: 3D People Surveillance Dataset 3DPes[14] [Baltieri *et al.* , 2011a] is a dataset designed mainly for person Re-ID and tracking. The dataset was captured by a multi-camera network of eight different cameras within a real surveillance scenario. Data were collected on different days. Since it is an outdoor dataset, it presents high variations of light conditions. Background models of the cameras are available and the 1012 snapshots of 200 persons are provided with silhouette masks and bounding box information. Some more details and benchmarking results on person Re-ID can be found in [Vezzani *et al.* , 2013; Baltieri *et al.* , 2015]. Some works have already proposed to use 3DPeS dataset in future work for **gait based Re-ID** [Kawai *et al.* , 2012].

–**KinectREID**: One of the few person Re-ID datasets collected using the Kinect sensor in an unconstrained environment is KinectREID [15] [Pala *et al.* , 2015]. The purpose of the dataset is to provide data to test and evaluate algorithms of **person re-identification** using features extracted from the Kinect sensor: anthropometry, **gait** and appearance of the clothes using both the skeleton features and RGB-D data. It is composed of many video sequences of 71 people, acquired indoor at various illumination conditions and various angles: three front, three behind and a side. Also, appearance variations *i.e.*, carrying backpacks, bags or other accessories were incorporated in the dataset.

–**Vislab KS20**: The KS20 Vislab Multi-view Kinect Skeleton dataset[16] [Nambiar *et al.*

[12] http://www.cvg.reading.ac.uk/PETS2009/a.html
[13] https://data.vision.ee.ethz.ch/cvl/aess/dataset/
[14] http://www.openvisor.org/3dpes.asp
[15] http://pralab.diee.unica.it/it/PersonReIdentification
[16] http://vislab.isr.ist.utl.pt/vislab_multiview_ks20/

Table 2.1: Characteristics of the main public datasets applicable to gait based Re-ID

Name	**#Camera**	**#Image resolution**	**#People**	**Scenario**	**Main application**
CASIA-datasetA	1(3 views)	352x240	20	Outdoor	Gait recognition
CASIA-datasetB	11	320x240	124	Indoor	Gait recognition
SOTON (large)	6	20	114	Indoor	Gait Recognition
USF	2	720x480	122	Outdoor	Gait recognition
SAIVT	8	704x576	150	Indoor	person recognition & Re-ID
AVAMVG	6	640 x 480	20	Indoor	Gait recognition
HDA dataset	13	2560x1600 (max)	85	Indoor	Person detection & Re-ID
i-LIDS (MCT)	5	576x704	119	Indoor	Person tracking
TUM-GAID	1(Kinect)	640x480	305	Indoor	Gait recognition
PETS2009	8	768x576	(NA)	Outdoor	Person detecion
PRID2011	2	64 x 128	245	Outdoor	Person Re-ID
ETHZ	1(stereo)	640 x 480	(NA)	Outdoor	Person detection & tracking
3DPeS	8	704x576	200	Outdoor	People Tracking and Re-ID
KinectREID	1(Kinect)	vary	71	Indoor	person Re-ID
Vislab KS20	1(Kinect)	NA	20	Indoor	person Re-ID

, 2017a] is a new dataset collected by the authors, in the context of long-term person re-identification. It comprises of multi-view Kinect skeleton (KS) data sequences collected from 20 walking subjects using Kinect v2. The major motivation behind the creation of this dataset was the lack of similar Kinect datasets consisting of people walking in different view-points, in order to assess the pose invariant long term Re-ID. Multiple walking sequences along five different directions i.e., Left lateral (LL at $\sim 0°$), Left diagonal (LD at $\sim 30°$), Frontal (F at $\sim 90°$), Right diagonal (RD at $\sim 130°$) and Right lateral (RL at $\sim 180°$) were collected. Altogether it has 300 skeleton image sequences collected from 20 subjects (3 video sequences per person in a particular viewpoint) in the aforementioned directions.

There are many other datasets individually available for Re-ID (*e.g.*, Viper [Gray *et al.* , 2007], CAVIAR4REID [Cheng *et al.* , 2011]) and gait analysis (*e.g.*, TUM-IITKGP Gait Database [Hofmann *et al.* , 2011], KY 4D Gait Database B [Iwashita *et al.* , 2014]). However, since we focus on the gait based Re-ID, we confine the search for the best among the set that serves our goal. Only those datasets made available for public access are described in this section; the other datasets proposed by single authors, as well as the local datasets are not considered. The details of the datasets above are summarized in Table 2.1.

(a) (b) (c)

Figure 2.3: Image samples from (a) PETS2009 (b) ETHZ and (c) 3DPeS datasets. All of them are outdoor surveillance scenarios, among which ETHZ has mobile cameras.

2.5.2 Anthropometry based Re-ID datasets:

In this section, we explain the datasets acquired using depth sensors which can provide the 3D body information, which is mostly insensitive to the appearance variations, and thus facilitate towards long-term person Re-ID.

–**RGB-D Person Re-identification Dataset:** The RGB-D dataset[17] [Barbosa *et al.* , 2012b] is a dataset proposed for person re-identification using depth information. The key objective of this dataset was to promote RGB-D re-identification research. It is composed of four different groups of data viz., Collaborative, Walking1, Walking2 and Backwards, collected using the Kinect upon 79 people. The acquisitions were performed in different days, with changes in appearances. Thus, this dataset is very well suited towards anthropometric based long-term person Re-iD analysis, leveraging RGB-D sensor information. Five synchronized data channels for each person were provided : 1) a set of 5 RGB images, 2) the foreground masks, 3) the skeletons, 4) the 3d mesh (ply), 5) the estimated floor.

–**BIWI RGBD-ID dataset**: Another similar RGB-D dataset proposed to perform long-term people re-identification from RGB-D cameras was BIWI RGBD-ID[18] [Munaro *et al.* , 2014b].It consists of video sequences of 50 different subjects, performing a certain routine of motions in front of a Kinect, such as a rotation around the vertical axis, several head movements and two walks towards the camera.The dataset includes synchronized RGB images, depth images, persons' segmentation maps and skeletal data, in addition to the ground plane coordinates. The videos are acquired at about 10fps.

–**IAS-Lab RGBD-ID Dataset**: Yet another long-term people re-identification dataset leveraging depth information is IAS-lab RGBD-ID[19] [Munaro *et al.* , 2014a].It contains 11 training and 22 testing sequences of 11 different people. The dataset includes synchronized RGB images, depth images, persons' segmentation maps and skeletal data, in addition to the ground plane coordinates. These videos have been acquired at about 30fps. For every subject, three sequences were collected making rotations and walkings, also in different clothings and at different room.

–**RobotPKU RGBD-ID dataset**: Another very recent dataset is RobotPKU RGBD-ID dataset[20] [Liu *et al.* , 2017]. The motivation behind this dataset was to perform more extensive experiments on a larger amount of data, and they collected RGB-D dataset called RobotPKU RGBD-ID Dataset with Kinect sensors. This dataset contains 180 video sequences of 90 person, and for each one the Still and Walking sequences were collected in two different rooms. This dataset includes RGB images, depth images, persons' segmentation maps and skeletal data.

[17] https://www.iit.it/research/lines/pattern-analysis-and-computer-vision/pavis-datasets/534-rgb-d-person-re-identification-dataset
[18] http://robotics.dei.unipd.it/reid/index.php/8-dataset/2-overview-biwi
[19] http://robotics.dei.unipd.it/reid/index.php/8-dataset/5-overview-iaslab
[20] https://github.com/lianghu56/RobotPKU-RGBD-ID-dataset

-**TVPR dataset**: The TVPR (Top View Person Re-identification)[21] [Liciotti *et al.* , 2016] dataset is a recently released dataset for top-view based person Re-ID. It stores depth frames (640x480) collected using Asus Xtion Pro Live in top-view configuration. This setup choice is primarily due to the reduction of occlusions and it has also the advantage of being privacy preserving, because faces are not recorded by the camera. The 100 people of TVPR were acquired in 23 registration session.

Also, the Kinect based datasets mentioned in the gait based Re-ID i.e.,KinectREID dataset,Vislab KS20 datasets are also available for anthropometry based Re-ID, since they comprise of multiple frames of walking image sequences. Hence, they are not repeatedly mentioned here. The summary of the anthropometry based long-term Re-ID datasets are presented in Table 2.2

Table 2.2: Characteristics of the main public datasets applicable to anthropometric based long-term Re-ID

Name	**#People**	**#View-points**	**#Frame rate**	**#Video sequence**	**Depth sensor**
RGB-D	79	Front, Rear	20	316 sequences	Kinect v1
BIWI RGBD-ID	50	Rotation, front	10	50 training and 56 testing sequences	Kinect v1
IAS-Lab RGBD-ID	11	Rotation, front	30	11 train and 22 test sequences	Kinect v1
RobotPKU RGBD-ID	90	Front, rear, rotation	30	180 video sequences	Kinect v1
TVPR	100	Top	30	23 registrations	Asus Xtion
KinectREID	71	Front, rear, lateral	30	483 video	Kinect v1
Vislab KS20	20	Left/ right lateral, left/ right diagonal, frontal	30	300 walking video sequences	Kinect v2

2.6 Re-ID performance evaluation metrics

Depending on the scenario and context of the application, the evaluation metrics employed in the Re-ID task may also vary accordingly [Vezzani *et al.* , 2013]. One noteworthy point is that for either appearance based or biometric based Re-ID, the evaluation metrics used are the same and therefore, in this section we analyse the performance evaluation metrics used for person Re-ID problem, in general. Here we present the different alternatives available for particular implementations of re-identification as either recognition or identification, as mentioned in Section 1.2.1. The major difference between recognition and identification is that the former has to point detected subject as a specific pre-defined unique subject (classification problem), whereas the latter has to establish the unique identity of the subject, without prior knowledge in an unsupervised manner (clustering problem).

2.6.1 Re-ID as recognition

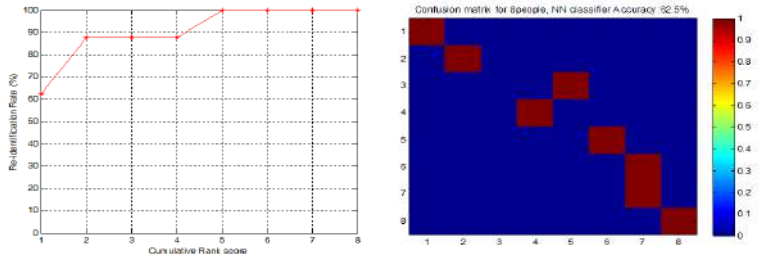
CMC curve: To evaluate the performance of Re-ID algorithms in closed-set scenarios, the cumulative matching characteristic (CMC) curve [Grother & Phillips, 2004; Phillips *et al.* ,

[21] http://vrai.dii.univpm.it/re-id-dataset

2000] is the most acclaimed and popular method of choice. The CMC curve shows how often, on average, the correct person ID is included in the best K matches against the training set, for each test image [Nambiar *et al.* , 2014b]. In other words, it represents the expectation of finding the correct match in the top K matches. Suppose a test sample (probe) is given rank-k, i.e., the subject is ranked in position k by the identification system. Since the identification rate is an estimate of the probability that a subject is identified correctly at least at rank-k, the identification rate is necessarily an increasing function of k and hence the name cumulative matching curve.

CMC measures how well the system ranks the identities in the enrolled database given the "unknown" probe image. Hence, the Re-ID task is considered as a recognition problem, with the assumption that only one sample class in the gallery corresponds to the query. As a result, the Re-ID output is given as a ranked list of gallery classes, based on some matching similarity to the query probe. Many often, this analysis is conducted by the forensics analysts during their investigation process, while evaluating a large set of image sequences against the probe [Vezzani *et al.* , 2013]. A comprehensive characterization of the CMC curve for evaluation or recognition problems is given in [Moon & Phillips, 2001], where it was originally proposed for the evaluation of face-recognition algorithms (on FERET image sets).

Fig.2.4(a) shows a typical CMC curve for the re-identification of a gait sample on a dataset of 8 people (an experiment conducted by the authors). Five persons were correctly re-identified in the first rank, which is represented by 62.5% in the cumulative rank 1. While considering top two matches, two more people were correctly re-identified (7 correct matches) thus improving the CMC curve towards 87.5%. And, this continues until all the people are properly re-identified. One key feature of CMC is that in a plot that includes all possible ranks (e.g., if the dataset has eight people, and the CMC goes through rank 8), the probability of identification is 100% at the highest rank (i.e., at rank 8). As a closed set identification process, it is showing the identification rate for the entire database. Many works in the area of gait analysis, as well as re-identification, employ CMC e.g., [Kale *et al.* , 2003; Farenzena *et al.* , 2010; Bialkowski *et al.* , 2012; Nambiar *et al.* , 2012, 2014b].

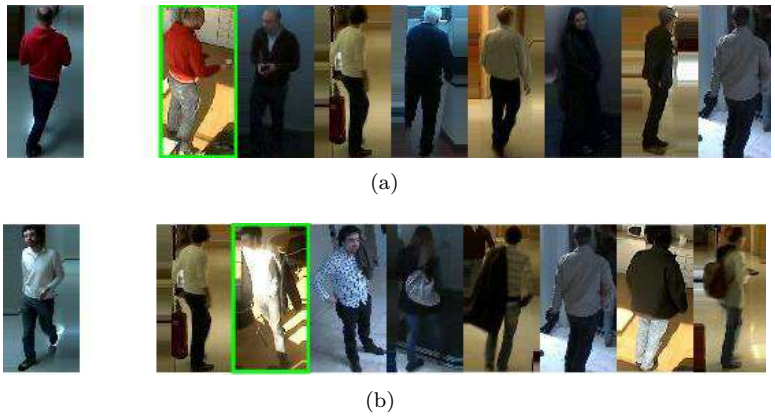


(a) (b)

Figure 2.4: Simple gait re-identification test with 8 persons with 2 sequences for training and 1 sequence for testing, per person. (a) Cumulative Matching Characteristic (CMC) curve showing the re-identification rate against the rank score (Rank1 accuracy= 62.5%);(b) Confusion Matrix showing the Re-ID accuracy.

Confusion Matrix: Another way of depicting the results of the re-identification is with the help of the confusion matrix. Each column of the matrix represents the instances in a predicted class, and each row represents the instances in an actual class (some examples are found in [Nambiar *et al.* , 2015; Middleton *et al.* , 2005; Bialkowski *et al.* , 2013; Aziz *et al.* , 2011b]). Each entry of the matrix contains the fraction of predicted cases classified as the actual ones. Diagonal terms express the accuracy of recognizing each class, and off-diagonal elements represent false classifications. The more "diagonal" is the matrix, the more accurate is the method. The confusion matrix inherits its name by the ability to inspect the cases of confusion, i.e. which classes are similar to the correct one and may lead to erroneous classifications. For instance, the confusion matrix of the Re-ID case conducted above for 8 people is depicted in Fig. 2.4(b). We could observe that Person ID 1, 2, 4, 7 and 8 were correctly classified leading to an accuracy of 62.5% while matching the predicted person ID vs actual person ID (i.e., person ID 3,5 and 6 were *confused*). This accuracy is same as the percentage of persons correctly re-identified in the first rank of CMC curve.

Performance visualization: A common qualitative method for representing the ranked list of gallery class against the test query is via visual representation. This is usually accomplished by plotting the ranked set of persons' bounding boxes (see Fig. 2.5). For the subjective analysis and the on-the-fly interpretation of the achieved result, this qualitative method is of great help. See some similar Re-ID results in [Vezzani *et al.* , 2013; Layne *et al.* , 2012].



(a)

(b)

Figure 2.5: Performance visualization: Examples of queries made in HDA person dataset reidentified by the methodology described in [Figueira *et al.* , 2013]. The probe query is shown at the left side and the top 8 results in the ranked Re-ID order is shown. The correct match is highlighted in green. (a) Test person is correctly identified in rank 1 (b) Test person correctly identified in rank 2.

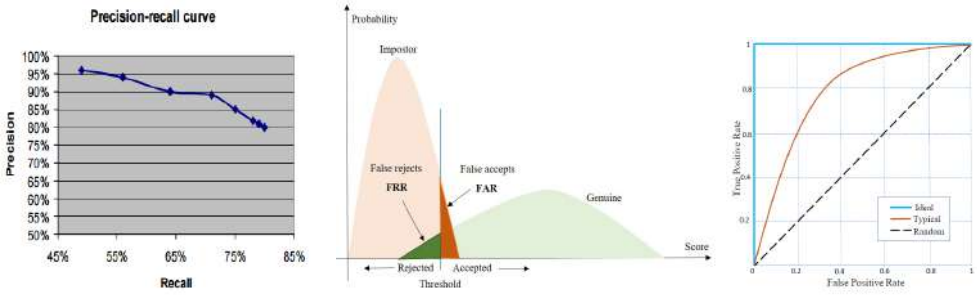
2.6.2 Re-ID as identification

Precision-Recall (P/R) statistics: In Re-ID scenarios related to identification/ verification, there may be instances to classify that are outside of the knowledge base. In this case, evaluation typically uses precision and recall (P/R) statistics. Another scenario where these metrics are suitable is, for instance in a shopping mall, where we want to track people in a camera network, but do not require their real identity. This scenario is similar to data clustering, where the correspondences among the set of people instances without prior knowledge have to

be established. Then, each cluster relates to an individual.

The performance evaluation of such a system works similar to a verification system, where it checks if two instances belong to the same person (1:1 biometric verification system). This analysis includes checking occurrences of false positives [22] and missed detections [23]. To fairly evaluate false positives, and appreciate the effect of missed detections, precision and recall statistics are more suitable than the CMC.

$$\begin{aligned} \text{Precision} &= \frac{\text{CorrectIdentifications}}{\text{TruePositiveDetections} + \text{FalsePositiveDetections}} \\ &= \frac{\text{CorrectIdentifications}}{\text{NumberofDetections}} \end{aligned} \tag{2.1}$$



(a) (b) (c)

Figure 2.6: Traditional curves used to evaluate the performance of Re-ID as a biometric identification/ verification system; (a) Precision-recall curve in the Re-ID experiment [Hamdoun *et al.* , 2008]; (b) FAR and FRR measures; (c) receiver operating characteristic (ROC) curve.

$$\begin{aligned} \text{Recall} &= \frac{\text{CorrectIdentifications}}{\text{TruePositiveDetections} + \text{MissedDetections}} \\ &= \frac{\text{CorrectIdentifications}}{\text{NumberofPersonAppearances}} \end{aligned} \tag{2.2}$$

An excerpt from [Hamdoun *et al.* , 2008] depicting a P/R curve is presented in Fig. 2.6 (a). Also some other Re-ID works employing PR curves are [Aziz *et al.* , 2011b; Figueira *et al.* , 2014; Satta *et al.* , 2014; Shi *et al.* , 2015]. Another interesting metric derived from precision-recall pair is F1-score (also F-score or F-measure), which acts as a measure of a test's accuracy. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0 (see some works in [Figueira *et al.* , 2014; Cancela *et al.* , 2014]).

FAR and FRR: There are some other standard biometric evaluation measures used in particular identification/ verification problems, viz., FAR, FRR, Receiver Operating Characteristics (ROC) curve and Decision Error Trade-off (DET) curve, that can also be used for re-identification. In biometric access control systems, it is common to use the trade-off between false acceptance rate (FAR) and false rejection rate (FRR). FAR is the percentage of

[22] False positive is an error in data reporting in which a test result improperly indicates presence of a condition, when in reality it is not

[23] Missed detection (false negative) an error in which a test result improperly indicates no presence of a condition, when in reality it is present

accepted non-genuine (impostor) individuals with the total acceptance made by the system. It is a measure of the likelihood that the system incorrectly accepts an access attempt by an unauthorized user. Similarly, false rejection rate (FRR) is the percentage of rejected genuine individuals compared to total rejects made by the system. It is the measure of the likelihood that the system will incorrectly reject an access attempt by an authorized user. Fig. 2.6 (b) shows a pictorial representation of false acceptance rate (FAR) and FRR.

An ideal human identification system requires the recognition performance with both FAR and FRR at zero level. Since it is impractical in real world applications, the threshold is determined by the type of application. For example, if Re-ID is to provide access control/ authentication purposes, the system prefers to keep FAR as low as possible (lower the access chance for impostors). In some other situations like forensic scenarios, the preference would be to reduce FRR, since we don't want to reject genuine individuals connected to the crime activity. FAR-FRR application to gait based person Re-ID was reported in Jungling et al. [Jungling & Arens, 2010], as well as in gait recognition [Wang *et al.* , 2003a].

Receiver Operating Characteristics: Receiver operating characteristic (Receiver Operating Characteristics (ROC)) curves are a well-accepted measure to express the performance of 1:1 matches. ROC curve plot is created by plotting the true positive rate (TPR- genuine users accepted) against the false positive rate (FPR- impostor users accepted) at various threshold settings (see Fig. 2.6 (c)). Since TPR is equivalent to sensitivity and FPR is equal to 1 - specificity, the ROC graph is sometimes called the sensitivity vs (1 - specificity) plot as well. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy [Zweig & Campbell, 1993].

An alternative to the ROC curve is the DET graph. A DET curve plots the false negative rate (missed detections) vs. the false positive rate (false alarm) on non-linearly x- and y-axes. Some instances of the applications of ROC and DET measures in gait analysis and Re-ID scenarios could be found in [Sivapalan *et al.* , 2012; Wang *et al.* , 2003a; Sivapalan *et al.* , 2011; Liao *et al.* , 2014].

2.6.3 Re-ID in forensics

In addition to the application in surveillance, Re-ID has found applications in forensics as well. As we have already mentioned the application of FAR and FRR evaluation metrics in forensic scenarios, there are also other standard measures commonly used for the same.

Likelihood Ratio: The likelihood ratio (LR) is a traditional measure in the forensics arena [Aitken & Taroni, 19995]. The LR is a standard measure of information that summarizes in a single number, the data support for a hypothesis [Perlin, 2010]. It is a good legal and scientific standing that underlies the credibility of forensic science in court by quantifying the belief in a hypothesis. Basically, LR is the ratio of two probabilities of the same event under different hypotheses. For two events, say A and B, the probability of A given B is true, divided by the probability of event A given B is false, is termed as a likelihood ratio [Vezzani *et al.* ,

2013].

$$\textbf{Likelihoodratio}(\textbf{LR}) = \frac{Pr(A|B)}{Pr(A|\neg B)} \tag{2.3}$$

In crime scenarios, $Pr(E|S)$ is the probability of the evidence (E) if the suspect is the source (s) of evidence, and $Pr(E|U)$ is the probability of the evidence if an unknown (U) is the source of evidence, then likelihood is calculated as follows

$$\textbf{Likelihoodratio}(\textbf{LR}) = \frac{Pr(E|S)}{Pr(E|U)} \tag{2.4}$$

Some applications have been reported in DNA analysis [Perlin, 2010] and in gait recognition [Muramatsu *et al.* , 2014].

Chapter 3

Anthropometry based Person Re-ID

Don't look for me in a human shape. I am inside your looking.

— Rumi

The direct computation of soft biometric features from video images is not trivial and existing methods rely on human manual measurements made on individual images [Reid & Nixon, 2011]. Instead, automated computer vision analysis methods have been more successful with features that are not interpretable by humans, like SIFT [Lowe, 2004], HOG [Dalal & Triggs, 2005], Shape Context [Belongie *et al.* , 2002] and others. These features, though useful in automated methods, are hard to reason about by humans and thus not suited for formulating verbal descriptions of search queries in databases. For instance, we would like to be able to search on a database for persons with large torso, thin neck, long head, etc. Thus, we propose a methodology to infer soft biometric person characteristics from their computer vision based descriptors, using regression analysis.

Obtaining a predictive model of soft biometric features from computer vision features involves several challenges and difficulties: (i) which computer vision features are more adequate; (ii) how to obtain the ground truth biometric features to train the model and; (iii) hwhich regression model is more suitable.

With respect to the first point, we propose the use of shape context (SC) features computed in the upper-torso part of the frontal human silhouettes, where we capture the human images from their video clips walking towards the camera. The upper torso region of the body presents less temporal variance with respect to arms and legs motions, thus producing more stable features. In addition to that, since person Re-Identification (Re-ID) is carried out in an uncontrolled environment, there are chances for clutters and other interacting objects making the lower body part occluded. In many indoor surveillance systems cameras are placed along corridors at high positions and tilted down, which makes the legs and lower torso occluded when persons are close to the camera. However, the head to chest region, unlike the waist and legs, maintain a relatively consistent shape through a broader range of walking frames. A real

world example (video sequence in the HDA dataset [1]) where this effect is clear is shown in Fig. 3.1.

Figure 3.1: Image showing the relevance of head-to-chest region for person re-identification: A forward walking sequence captured in our HDA dataset highlights the visibility of head-to-chest region in most of the frames, while the other parts are occluded.

The use of a silhouette based feature is motivated by the fact that it is less sensitive to the color and texture of the inner region of person's images and thus making itself a better candidate towards long term based person Re-ID. Furthermore, the shape context (SC) feature computes the density of boundary points at various distances and angles. As such, it more directly encodes soft biometric traits such as lengths, curvatures and size ratios in the human body. We explore this idea with the goal of recovering the soft biometric features encrypted in the SC descriptors of body silhouettes using regression methods.

The second challenge is the availability of ground truth biometric features to train the regression model. It is not easy to model this in a real environment due to the necessity of a range of variations of discriminative biometric features in relatively large population. Also, it is laborious to annotate the human biometrics manually on real data. In order to tackle this issue, we used Synthetic Avatars in a Virtual reality platform. In contrast to [Reid & Nixon, 2011], where the training set was generated by manual annotations done on real imagery by a large number of human annotators, here we avoid such a troublesome training phase by generating the ground truth with the help of modern computer graphics technology.

We leverage on the ability to simulate thousands of variations in biometrics on avatars according to our choice for two purposes. First we conduct a baseline study to verify the impact of our descriptor for Re-ID, since the simulated avatars provide flawless silhouette images. Second, we model the regression between computer vision based features SC and human interpretable Biometric Features (BF) and thus bridge the gap between the human and machine interpretations of human body shape. Thus we present a novel automatic person retrieval system which could work in dual mode (viz., multimedia mode or human query mode) depending on the test data.

For obtaining some geometric features of the head to chest region, distinguishable from person to person, some measurable metrics which vary significantly within the population should be chosen. The measurement and study of such features and their variation is the domain of anthropometry. An anthropometric survey (ANSUR) was conducted by the U.S. military in 1988 upon more than 150 anthropometric dimensions, measured from 9000 soldiers [Gordon

[1] http://vislab.isr.ist.utl.pt/hda-dataset/

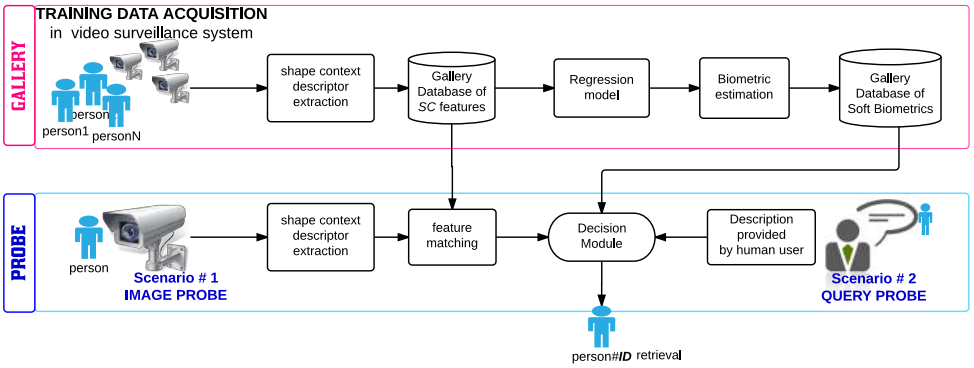


Figure 3.2: The scheme presents the framework of our human identification system. The probe data can be either the images/videos of the subject to identify (*Scenario#1*), or a description of the subject provided by a human operator such as eyewitness statement in a criminal scene (*Scenario#2*).

et al. , 1989]. A statistical summary of those standard biometric features related to the upper torso regions is provided in Table 3.1. In our study, we consider some of those key biometric features described here.

Table 3.1: A summary of anthropometric data taken from [Gordon *et al.* , 1989] relevant in the upper torso region. These statistical summaries reveals significant variation in the head and chest measures (all measurements are in centimeters). The features shown in bold letters are some of the soft biometric cues used in our study

Measurement Name	Mean	Standard Deviation	Min	Max
Biacromial Breadth	39.70	1.80	33.0	45.10
Bideltoid Breadth	49.18	2.59	41.0	59.3
Head width	15.51	0.60	13.6	17.7
Head circumference	56.77	1.54	51.4	62.7
Head Length	20.02	0.72	17.6	22.6
Chest Breadth	32.15	2.55	25.70	42.20
Thelion length	27.24	1.81	22.2	34.2
Neck circumference	37.96	1.97	31.6	47.0
Shoulder circumference	117.52	6.04	96.6	142.4
Shoulder-elbow length	36.9	1.79	29.7	44.6
Shoulder length	15.05	1.10	11.4	18.5

The third challenge is to find the best regression model, indicating the relationship between biometric features and image features. Concerning this, we employ both choices - linear and non-linear regression schemes under the Support Vector Regression (SVR) framework. i.e., by employing SVR as the instance of regression, with linear as well as nonlinear kernels. We also conduct an extensive study on the selection of meta parameters for each of these kernels and finally, we devise the best among them for the design of the regression block in our system. A detailed explanation of our experimental dataset, SVM regression, the choice of basis, meta parameters and the cross validation strategies are provided in section 3.2.2.

3.1 System Architecture

The general framework of our Re-ID system is presented in Fig. 3.2. Our scheme is designed to work in two different modes depending on application scenarios. In the first scenario (***Scenario#1***), the test images/ videos of the subject to identify are provided, whereas in the second mode (***Scenario#2***) the probe is solely a verbal query or a description of the subject provided by a human operator. The former scenario mainly concerns the use of multimedia content for re-identification of a suspect in a video surveillance network by extracting his feature descriptors and matching with gallery database. The latter scenario instead does not require any multimedia content but exploits the eye-witness description of the suspect related to biometrics cues such as short neck, large chest etc. We note here that these human descriptions are analogous to *human compliant labeling* referred by [Dantcheva *et al.* , 2010] or *semantic annotation* referred by [Samangooei *et al.* , 2008]. Rather than re-identifying, this mode is more pertinent towards categorizing the population based on their respective human compliant traits and thus retrieving their identifier (*#ID*). Since many people could have similar semantic labels resulting in subject interference, grouping them into classes with similar traits could be the best technique to tackle this issue. This is a kind of pruning method, which normally the security people do manually on receiving the human queries; we do it here automatically.

The general framework of the system is presented in Fig. 3.2. In the training phase, human video footage is acquired in a video surveillance system and stored in gallery. The camera network is connected to the SC descriptor module, where the acquired persons' SC features are extracted. These extracted SC features are stored in a gallery database for later use. The database is accessible by the feature matching module, which has the purpose to compare the feature descriptor of the person we want to identify (i.e., the probe) with the ones stored in the database. In ***Scenario#1***, when a new image frame of the person is acquired, his SC descriptor is extracted and compared with those in the gallery set. In the decision module, based on the matching similarity measurements, the most similar person *ID* in the training set is retrieved thus facilitating the system for person Re-ID.

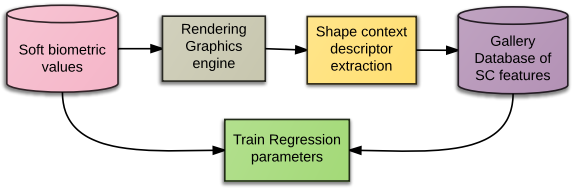


Figure 3.3: Framework for training the regression model

Another major module of the system in the training phase is a regression block connected to the database of SC features. It divulges the relation of SC descriptors with soft biometrics, and it estimates BF corresponding to each sample. These estimated biometric values are stored in a gallery database of biometrics, which is connected to decision module. The decision module analyses these biometric data and carries out a statistical analysis among the population. In ***Scenario#2***, when the probe input in terms of human query enters in the decision module,

it will examine the statistical profile of the population and retrieve the category of suspect. As a result, all the person *ID*s' in the suspected category, as well as the tentative ranked list of suspect are published.

To learn the regression module, we require a vast and vivid benchmarking dataset to substantiate the mapping between SC feature space and biometric space. For that purpose, we generate avatars in virtual reality and carry out regression analysis. When a new SC feature is received, the corresponding output biometric values are estimated based on this regression model. In addition to that, a virtual reality population is also employed to verify our methodology towards person re-identification and to compare with the counterpart experiment in real scenario.

3.2 Methodology

This section describes the main ingredients of the proposed approach. Basically it comprises: (i) computation of the SC features from the images containing the head and torso, (ii) matching the SC between two head-torso silhouettes and (iii) the statistical regression analysis between SC and the space of BF.

3.2.1 Feature extraction

Shape Context

The original idea of shape context (SC) was described in the paper of [Belongie *et al.* , 2002]. In order to achieve the shape similarity or the shape distance, they introduced a new descriptor called *shape context*, which measures the distribution of points in a shape relative to each point in that shape.

Fig. 3.4 depicts the method of obtaining SC descriptors. The silhouette of an object is sampled at N discrete points along the contours, $P= (p_1,p_2,....,p_N)$. For a point p_i, a coarse histogram h_i of the relative co-ordinates of the remaining N-1 points is identified and is termed as the SC of p_i:

$$h_i(k) = \#\{q \neq p_i : (q - p_i) \in bin(k)\} \tag{3.1}$$

Thus, a compact and highly discriminative descriptor is computed as the distribution over these relative positions. A uniform binning scheme in log-polar space is adopted making the descriptor much more sensitive to nearby sample points than to those farther away. As shown in Fig. 3.4(c), we use 12 equally spaced angle bins and 5 equally spaced log-radius bins, altogether making the dimension of the SC as 60. In contrast to the closed object shapes proposed in the original work, we apply the re-sampling on the open shape of the silhouette, i.e., we don't consider the cropping line in the chest. To represent each silhouette (Fig. 3.4(a)) we used 40 points uniformly sampled from the Canny edges (Fig. 3.4(b)). Then we flattened and concatenated the complete set of 40 sample point SC each with 60 dimensions, thus producing a SC histogram of dimension 2400.

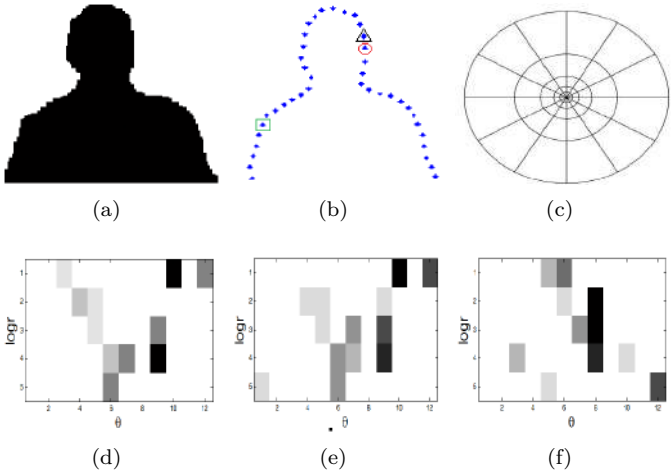


Figure 3.4: Shape context computation. (a) Silhouette of upper human body part (b) Sampled edge points of the silhouette shape (c) Diagram of log-polar histogram bins used in computing the SC. We have used five bins for *logr* and 12 bins for θ. (d,e,f) corresponds to the SC for reference samples marked by △, ◯ and □. Visual similarity of the SC for nearby points △, ◯ is pretty obvious whereas the SC of the □ point, is quite different. (Note: Dark= large value)

Matching Shape Context

In order to compare two different shapes we must define a similarity metric. To mitigate problems of misalignments of the silhouettes' sampling points due to discretization, a previous alignment step is necessary. Two criteria are to be met while matching SC features: (1) corresponding points should have very similar descriptors, and (2) the correspondences should be unique.

First criteria is handled via cost matching technique. Let C_{ij} denote the cost of matching two sample points p_i and q_j in two different shapes, by means of χ^2 test statistics

$$C_{ij} = C(p_i, q_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k) - h_j(k)]^2}{[h_i(k) + h_j(k)]} \tag{3.2}$$

where, $h_i(k)$ and $h_j(k)$ denote the K-bin normalized histogram at p_i and q_j, respectively. Given the set of costs C_{ij} between all pairs of points, the uniqueness criterion is addressed as follows. To match two shape contours say, P and Q, we minimize the total cost of matching

$$H(\pi) = \sum_i C(p_i, q_{\pi(i)}) \tag{3.3}$$

subject to the constraint that the matching is one-to-one, i.e., π is a permutation. This is an instance of the square assignment (or weighted bipartite matching) problem. In our experiments, we make use of the Hungarian algorithm [Harold, 1955].

3.2.2 Regression

Regression analysis is a statistical process for estimating the relationships among variables. The technique is widely used in machine learning for prediction and forecasting. It is also

helpful to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. Here, the relationship between SC and BF features is learned via Regression analysis. We exploit Support Vector Regression (SVR) as an instance of regression, and we experiment with linear and non-linear kernels.

Referring to Fig. 3.5, we can understand that a semantic bridge between machine derived computer vision features and the human interpretable biometrics features is realised via a Regression module. The rationale is that, by dint of this regression module, the relationship of verbal descriptions and their corresponding images could be learned a priori. And, hence, when a verbal query is made in the biometric space, the system can easily relate it to with its counterpart in the image space.

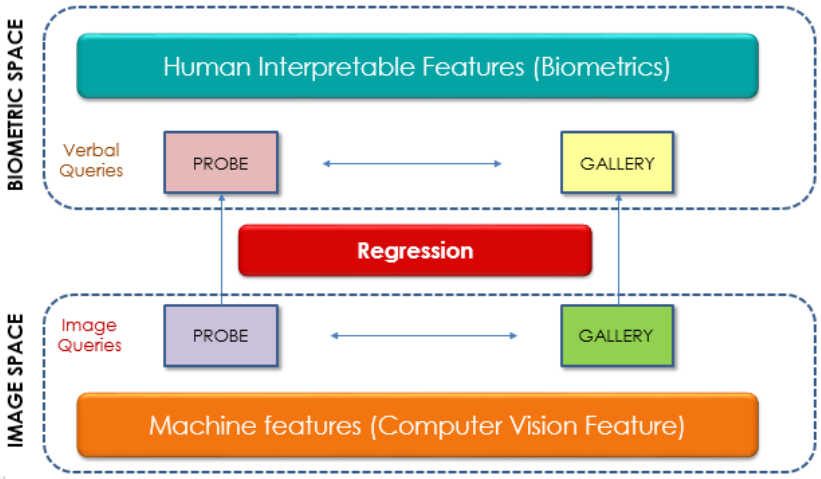


Figure 3.5: The functional diagram of the Regression scheme in our proposed architecture.

Support Vector Regression

Support Vector Machines [Cortes & Vapnik, 1995] can be applied not only to classification problems but also to the case of regression [Smola & Schölkopf, 1998], [Chapelle & Vapnik, 1999]. Analogous to Support Vector Classification, the produced model depends only on a subset of the training data, in SVR as well. In ϵ-SVR [Vapnik, 1995], the goal is to find a function *f(x)* that has at most ϵ deviation from the desired targets y_i for all the training data. In other words, ignore errors as long as they are less than ϵ, but will not accept any deviation larger than this.

Training the original SVR as per [Smola & Schölkopf, 2004] is mentioned below. Suppose the training data $\{(x_1, y_1), ...(x_n, y_n)\} \subset \mathcal{X} \times \mathbb{R}$ where, $\mathcal{X}$ denotes the space of the input patterns. (e.g. $\mathcal{X} = \mathbb{R}^d$). More specifically, $(x_1, , ...x_n)$ is the set of independent variables and $(y_1, , ...y_n)$ is the set of corresponding dependent variables in the training set. Consider linear functions *f*, taking the form

$$f(x) =< w, x > +b, \;\; w \in \mathcal{X}, b \in \mathbb{R} \tag{3.4}$$

where $< ., . >$ denotes the dot product in $\mathcal{X}$ with w being the normal vector to the separating hyper plane, and b being the bias term. Solving this convex optimization problem according to [Vapnik, 1995] yields the formulation below. In order to cope with the infeasibile constraints of the optimization problem, slack variables ξ_i, ξ_i^* are introduced.

$$\begin{aligned} &minimize \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \\ &s.t. \begin{cases} y_i - < w, x_i > -b \leq \epsilon + \xi_i \\ < w, x_i > +b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{3.5}$$

The constant $C > 0$ determines the trade-off between the flatness of such a function f and the amount up to which deviations larger than ϵ are tolerated. This corresponds to dealing with a so called ϵ-insensitive loss function $|\xi|_\epsilon$ described by

$$|\xi|_\epsilon = \begin{cases} 0, & \text{if } |\xi| \leq \epsilon \\ |\xi| - \epsilon, & \text{otherwise} \end{cases} \tag{3.6}$$

To extend this towards nonlinear functions, the main strategy is dual formulation. Hence, the optimization problem could be transformed into a dual problem and its solution is given by:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x_j) + b, \quad s.t., \begin{cases} 0 \leq \alpha_i^* \leq C, \\ 0 \leq \alpha_i \leq C \end{cases} \tag{3.7}$$

where, α_i, α_i^* are the dual variables and $K(x_i, x_j)$ is the Kernel function.

The performance (estimation accuracy) of Support Vector Regression (SVR) depends on a good setting of meta parameters. The problem of optimal parameter selection is further complicated by the fact that SVM model depends on those kernel parameters.

Choice of Basis

We tested two kinds of regression bases: *(i)* Linear basis

$$K(x_i, x_j) \equiv < x_\mathrm{i}, x_\mathrm{j} > \tag{3.8}$$

which implies that the regressor is linear with respect to the input vector; *(ii)* Radial basis kernel as an instance of nonlinear basis, where

$$K(x_i, x_j) \equiv exp(-\gamma||x_\mathrm{i} - x_\mathrm{j}||^2), \ \gamma = \frac{1}{2\sigma^2} \tag{3.9}$$

The kernel trick avoids the explicit mapping that is needed to get linear learning algorithms to learn a nonlinear function; instead we use Kernel functions. We apply various regression analysis, on both of these Linear SVR and Nonlinear SVR, also by tuning their meta parameters. The forthcoming section explains the process in detail.

Kernel Parameters

We leverage the ϵ-SVR package from LIBSVM [2] library for SVM regression analysis. The meta-parameters to set are the cost, the kernel width, and the width of the insensitive zone, respectively C, γ and ϵ in equations (3.5), (3.9) and (3.6).

Parameter C determines the tradeoff between the model complexity (flatness) and the degree to which deviations larger than ϵ are tolerated in optimization formulation. For example, if C is too large (infinity), then the objective is to minimize the empirical risk only, without regard to model complexity part in the optimization formulation.

Parameter ϵ controls the width of the ϵ-insensitive zone, used to fit the training data. The value of ϵ can affect the number of support vectors used to construct the regression function. For lower ϵ, fewer support vectors are selected. On the other hand, bigger ϵ values results in more 'flat' estimates. Hence, both C and ϵ values affect model complexity.

Another important RBF parameter is γ, which will determine the width of the Gaussian kernel used i.e., $\gamma = \frac{1}{2\sigma^2}$. Increasing γ will increase the curvature of the fitting curve. Intuitively, γ defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.

Grid search and Cross validation

It is not known beforehand which values of C and γ are best for our problem. Consequently, we carry out some kind of model selection (parameter search) by means of exhaustive "grid search". The ultimate goal is to identify good (C,γ) so that the SVR can accurately predict the unknown data (testing data). In "grid search", different pairs of values are tried and the one with the best cross-validation accuracy will be chosen. We applied the baseline approach of trying exponentially growing sequences of C and γ. (C =2^{-15},2^{-13}.....,2^{15}; γ=2^{-15},2^{-13}.....,2^{15}). For each pair, we measure the prediction error (Mean Squared Error (MSE)), and the lowest MSE corresponds to the best result. MSE of a predicted value $\hat{y}$ of a regression's dependent variable y is computed for n different samples as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{3.10}$$

Consequently, the corresponding C and γ values which produced the least MSE are considered as the best meta parameters.

Also, in order to overcome the problem of overfitting, we carry out cross-validation. Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent dataset. One round of cross-validation involves partitioning

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. In our experiment, we analyzed two different modes of cross validation:

(a) K-fold cross validation: We first divide the training set into k subsets of equal size. Sequentially, one subset is tested using the regressor trained on the remaining k-1 subsets. This rotation estimation will go on k times, and finally, the prediction errors MSE over k folds will be averaged to produce a single estimation.

(b) Stratified K-fold cross validation: In stratified k-fold cross validation, the folds are selected such that each set contains approximately the same percentage of samples of each target class as the complete set. "Stratified" cross-validation is a simple variant of classical k fold cross-validation. When we do the initial division into k parts, we ensure that each fold has got approximately the correct proportion of each of the class samples. It basically makes sure that we choose a division that has approximately the right representation of class values in each of the folds. It helps reduce the variance in the estimate a little bit more.

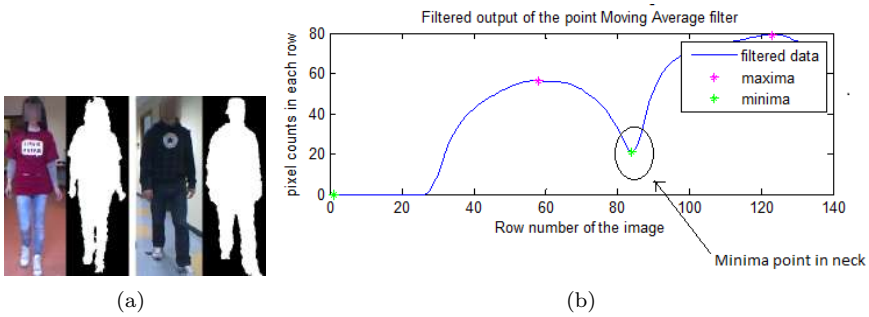
After the cross validation is done, we will get a single estimation of the measure of fit viz., the average *MSE_Train*[3] and the corresponding meta parameters. Based on this model, we train the whole system so that whenever the test data enters, it will estimate the output variables.

3.3 Experimental setup

We conducted the experiments in two modes (refer to our Re-ID system in Fig. 3.2). First, we carry out experiments in ***Scenario#1*** to study the feasibility of upper torso SC for person Re-ID. Initially we conduct a study with an existing person Re-ID dataset. However, the real world scenario is prone to segmentation noise. Thus, in order to validate our system in a noise free environment, we conducted our second experiment in a simulator platform, using virtual reality avatars. We simulated custom avatars corresponding to the humans in the real world, and conducted our experiments on them as well.

The second mode of experiments is done in ***Scenario#2***. Here, we explore the relationship between SC descriptors and BF by means of regression. Thus, we bridge the gap between human and the machine definition of biometrics with aid of computer vision and machine learning techniques. One noteworthy aspect is that in both experimental modes, the system doesn't require the co-operation of the subject as in hard-biometric data acquisition, thus making this soft biometric system very suitable for surveillance applications, where such cooperation is hard to achieve.

[3]Note that, we define *MSE_Train* and *MSE_Test* as the MSE obtained in both the training and test sets.



(a) (b)

Figure 3.6: (a) Sample images and corresponding silhouettes in our real-world experiment. (b) Pixel count value curve facilitating the automatic cropping of the upper body region by observing the minima point in the neck.

Scenario#1

We conducted a pilot study in the real world, where we incorporate the human silhouettes captured using KINECT camera in RGB-D person re-identification dataset [4] [Barbosa *et al.* , 2012a]. Along with each human image, corresponding human silhouette information is also provided. An example of the dataset is seen in Fig. 3.6(a). For our studies, we made use of their 'walking1' and 'walking2' categories, where we can obtain frontal appearance of the walking people. As a case study we only used 20 people, each one with 4 samples. There are images with different backgrounds, and the same person in different dressing, thus making the dataset very suitable to study the impact of our methodology in long-term person identification based on shape of the silhouette.

Since we are interested only in the upper torso region, we split the body into two parts and select only the region of interest i.e., upper part. Then we try to localize the neck location, which could be acted as key point. As seen in Fig. 3.6(b), the pixel count along the row of the image is plotted against the row number, which depicts the variation of the silhouette's thickness. We apply moving average filter to smooth out the fluctuations in the data curve. A key point corresponding to the neck is found by searching the minimum in the curve. Next, a standard amount of height equal to head to neck, is added towards bottom onto the chest region from the neck point in order to define the crop line in the chest. Afterwards, we normalize the height and rescale the width of the cropped region, maintaining the aspect ratio.

Prior to conducting the experiment of person re-identification, we had to apply some initial pre-processing steps to address the problem of silhouette imperfection mostly occurring due to segmentation errors and pixel noise. To get rid of the void spaces in silhouettes and to attain data quality, we applied morphological operations such as dilation followed by erosion. Afterwards, while the silhouettes are ready for our experiment, we equally split the dataset into half. Former set is the training set and the latter is the test set. In real world scenario, out of 80 sample images, we have 40 samples in both gallery and probe, i.e., 2 samples per 20 different persons are made available in both training and test set. Afterwards, the SC descriptor for each silhouette in the gallery is calculated. When the test set is provided, its matching cost

[4] http://www.iit.it/en/datasets-and-code/datasets/rgbdid.html

towards each of the 40 gallery samples is found using the Hungarian method. Then, each test sample will search for the minimal cost between itself and the gallery descriptors. The gallery sample with minimal cost (i.e., maximum similarity) and is selected as the best matching.

Custom avatars for re-identification: In the previous section we discussed about the experiments conducted in real database. In this section we evaluate the influence of the noise of the segmentation while extracting the head-to-torso region. To perform this study, we replicate the real dataset with virtual reality avatars leveraging computer graphics tool (the game engine *Unity3D®*) that allow us to render and manipulate the shape of synthetic humans.

We used some standard avatar packages viz. Male character pack and female character pack from Mixamo 3D[5] character animation service and Character Pack 02 from Animation arts Creative GmbH[6]. We modelled the custom avatars as close as the corresponding human instances by matching their shape traits and incorporating the posture and inclination of shoulders. Samples of the real human instances and their corresponding custom avatar models are illustrated in Fig. 3.7. After generating the custom avatars, we executed walking animations of these avatars and captured random 4 frames for each person which resembled the video surveillance image acquisition. Thus our virtual reality dataset also consisted of 80 synthetic samples corresponding to the 20 human instances in real world experiment. Then, we split them into gallery and probe and conduct descriptor matching in the same way conducted for real world dataset.

(a) #female1 (b) #female2 (c) #male1 (d) #male2

Figure 3.7: Sample instances of custom virtual avatars simulated corresponding to the real world dataset.

Scenario#2

Generic Avatars for regression: Albeit we simulated *Custom* avatars in our previous experimental setup, the dataset was limited in terms of variability of biometric features since only 20 human instances were generated in the simulator. In order to compute the regression model between SC features and BF, this was not enough to represent variation range of the real human population. Thus, we introduced a more global avatar set called as *Generic* avatars, by imposing larger variabilities as observed in the human population. Such a generic population is preferred over custom avatars for modelling the regression map, since it covers wider ranges of features. By incorporating extremal shapes, the generic dataset provides a higher

[5] https://www.assetstore.unity3d.com/en/#!/publisher/150
[6] https://www.assetstore.unity3d.com/en/#!/publisher/6659

signal-to-noise ratio[7] available for regression analysis.

Figure 3.8: Six standard avatars used in the synthetic platform for the generation of large dataset by changing the biometric features. We make use of only the upper-torso region including head, shoulder and chest.

Again we exploit the graphics engine *Unity3D*® to simulate the multiple avatars in virtual reality. Here we used 6 standard avatars viz. Male character pack and female character pack (shown in Fig. 3.8) from Mixamo 3D character animation service, as the baseline avatars. The default avatars models available in the package were considered as standard models, in which we assumed a unitary scale factor of each biometric measurement (see Fig. 3.9(a)). Afterwards, we generated the other avatars by imposing variations to the biometric features with respect to this standard model in the *Unity3D*® platform. The scale parameters of the avatar examples are defined by analysing the variability in real world human population. Here are the biometric features we employ in our experiment:

- Neckness (N) : length of the neck
- Chestsize (C) : horizontal distance between the lateral margins of the upper torso
- Bodysize (B) : Overall body size
- Headlength (HL) : maximum vertical length of the head
- Headwidth (HW) : maximum horizontal width of the head

Table 3.2: Chart showing the soft biometric scale factors for the simulated avatar versions in Figure 3.9. Values highlighted in bold characters in each row represents the modification imposed for that particular avatar.

Avatar Index	**Neckness (*N*)**	**Chestsize (*C*)**	**Bodysize (*B*)**	**Headlength (*HL*)**	**Headwidth (*HW*)**	**Human description label**
(a)	100%	100%	100%	100%	100%	Standard
(b)	**200%**	100%	100%	100%	100%	Large neck
(c)	**300%**	100%	100%	100%	100%	Very large neck
(d)	100%	**200%**	100%	100%	100%	Large chest
(e)	100%	**300%**	100%	100%	100%	Very large chest
(f)	100%	100%	**50%**	100%	100%	Thin body
(g)	100%	100%	**200%**	100%	100%	Fat body
(h)	100%	100%	100%	**125%**	100%	Long head
(i)	100%	100%	100%	100%	**125%**	Wide head

Table 3.2 shows the soft biometric parametrization imposed for simulating generic avatar population. Each value in the table corresponds to the scale applied to the standard model counterpart of that anthropometric measurement. We alter the biometric values one at a time by keeping other features intact. Thus, as per mentioned in the table, we can have 8 different

[7]Considering the noise in the SC features as constant (discretization noise), a higher variability in the range of the features (signal) will result in a better signal-to-noise ratio that will improve the quality of the regression model.

modified avatar models generated out of the standard avatar, by altering each biometric feature individually. Fig. 3.9 shows an example of the different virtual avatar samples generated out of a single basic standard avatar.

Fig. 3.9(a) shows a standard avatar where all the parameters are normalised (100%). Fig. 3.9(b) and Fig. 3.9(c) correspond to 200% and 300% Neckness, which intuitively means those models' neck length is twice and thrice longer compared to the standard one. Fig. 3.9(d) and Fig. 3.9(e) illustrate the 200% and 300% chest sized avatars respectively. Thin body size and Fat body are generated in Fig. 3.9(f) and Fig. 3.9(g) by setting scaling the body size parameter by 50% and 200% respectively. The last two avatars concentrate on the geometric parameters of head, by increasing 25% horizontally (head width) and 25% vertically (head length). Thus, we managed to generate an approximate variation of biometric features in synthetic population as observed in the human population. The idea was to be able to cover the range of variability as much as possible with the least number of examples. This way we could enhance the signal-to-noise ratio of the regression analysis.

Altogether 9 variations were generated out of each of 6 standard avatar. Then, we executed walking animations and captured random 4 frames for each person which resembled the video surveillance image acquisition. Thus our *Generic* avatar dataset consisted of 216 images.

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 3.9: The nine variations of biometrics simulated in the generic avatars. Only the upper torso region is shown since it is the region of our interest. Please refer to the Table 3.2 for measurement details.

Dataset for Regression

Suppose the regression is carried out from an input space of dimension $\mathbb{R}^p$ to an output space of dimension $\mathbb{R}$. Each element in the input space is a feature vector of size $p \times 1$. i.e. $\mathbf{x} = \left[x^1, \cdots, x^p\right]^T$. We collect n such samples and represent them as a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ as follows:

$$\mathbf{X} = \left[\mathbf{x}_1 \mid \cdots \mid \mathbf{x}_n\right]^T \tag{3.11}$$

Each row in the $\mathbf{X}$ matrix represents a feature vector corresponding to the n'th sample in the dataset. We collect the response variables y_i corresponding to each input sample $\mathbf{x}_i$ and represent them as a vector $\mathbf{y} \in \mathbb{R}^{n \times 1}$, as follows:

$$\mathbf{y} = \left[y_1, y_2, y_3 \ldots, y_n\right]^T \tag{3.12}$$

In our case, $\mathbf{X}$ contains the input SC descriptors and $\mathbf{y}$ holds the Biometrics values of the simulated avatars. We have 216 avatar samples corresponding to 4 different views of each of

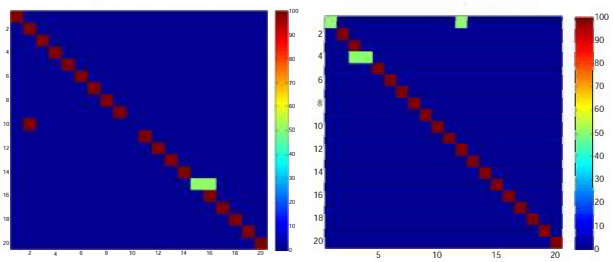
the 54 different shapes. The Shape Descriptors are composed of 40 points across the edge of the upper torso silhouette, each with 60D shape context descriptor and thus producing a 2400 dimensional feature vector corresponding to a person, i.e., input matrix $\mathbf{X}$ is of dimension $\mathbb{R}^{216\times2400}$. The output biometric consists of 5 biometrics say, *BF= (N, C, B, HL, HW)*. In our experiment, we conduct regression analysis individually for each of the biometrics in the set *BF*. More specifically, $\mathbf{y}$ in equation (3.12) will be a vector of dimension 216 containing a given biometric feature for all the avatars. Thus, each regression analysis will be from $\mathbb{R}^{216\times2400}$ matrix to $\mathbb{R}^{216}$ vector.

For the learning phase of the regression, we needed to have a benchmark dataset, with the corresponding image descriptor-biometric pair information. In order to employ this task, we used the same *Generic avatars*, we used in *Scenario#2*. Altogether 9 variations were generated out of each of the 6 standard avatars. Then, we executed walking animations and captured random 4 frames for each person which resembled the video surveillance image acquisition. Thus our *Generic* avatar dataset consisted of 216 images.

3.4 Results & Discussion

3.4.1 Person re-identification using Shape Context

Regarding both the experiments in ***Scenario#1*** (refer Fig. 3.2 and Section 3.3), the goal is to retrieve the most similar person in the gallery set for a given test person, by matching its SC descriptor with those in the gallery. Or in other words, when the probe imagery of the suspect is provided, its shape similarity with all the other training images in the gallery is measured by bipartite graph matching technique on SC features and the person re-identification is carried out.



(a) (b)

Figure 3.10: (a)Confusion Matrix showing a re-identification accuracy of 92.5% among the 20 humans in the real world scenario.(b)and 95% among the 20 custom avatars simulated corresponding to the instances in the real human dataset.

We depict the result of re-identification with the help of confusion matrix. Our results of person Re-ID is illustrated in Fig. 3.10. The first result in Fig. 3.10 (a) is the *Confusion Matrix* corresponding to our study with 20 real world instances and showing a re-identification accuracy of 92.5%. Fig. 3.10(b) is the counterpart *Confusion Matrix* in virtual setup with 20

custom avatars, and it achieved 95% accuracy in Re-ID. In both cases, we could observe high performance of our proposed SC algorithm to re-identify people. This accentuates the feasibility of utilizing shape as an effective soft-biometric cue in re-identification scenarios. Moreover, by conducting the comparative study in virtual setup, we could observe the influence of segmentation noise in reducing the Re-ID rate in the real world scenario.

3.4.2 Regressor performance

In this experiment, we analyse the regressor performance, to test linear and nonlinear models on the ability to predict biometric features from image data. As we explained earlier, we conducted experiments using the database of 54 avatars, with 4 samples each. Then, the regression analysis is conducted from the input space of $\mathbb{R}^{2400}$ to output space of $\mathbb{R}$. Among the output biometrics to be estimated say, *BF= (N, C, B, HL, HW)*, we perform regression analysis individually for each of them i.e., we regress the scalar estimate of each biometric from a 2400-D shape context vector. In our experiments, we selected 2 random avatars, each with 4 samples (total 8 samples), as the test set and the remaining 52 people, each with 4 samples as the training set (total 208 samples).

We conducted 6 different experiments on our data, over different kernels as well as different cross validation schemes. Out of these experiments, we report the Mean Squared Error (MSE) viz., *MSE_Train* and *MSE_Test* in both the training and test sets, as well as the best meta parameters (the ones leading to the least *MSE_Train*).

Table 3.3 summarizes the test and train set performances of the various regression methods studied on a single biometric feature (Neckness). Linear and kernelized basis versions were tested with different cross validation schemes, at manual and optimal regularizer settings. *MSE_Train* corresponds to the MSE obtained for the training set obtained via cross validation, and the *MSE_Test* is the the MSE obtained for the test set. In the default parameter setting, the default meta parameters are activated (C=1, γ=1/num_features, ϵ=0.1), whereas in exhaustive grid search, the optimal values of meta parameters are selected as the pair of (C,γ) producing the least *MSE_Train* in the training set. A sample grid search selection of optimal meta parameters for *Expt.5* (in Table 3.3) is depicted in Fig. 3.11(a).

In order to verify the repeatability/consistency of the measure of fit, we executed 10 runs of random trials (with Cross validation of 2 fold) for the same biometric. The boxplot representation of the variability of regression performance in terms of Mean Squared Error for both train and test sets are shown in Fig. 3.11(b) and Fig. 3.11(c), respectively.

Next, we try to extend the case studies conducted on a single biometric feature (Neckness), over all the 5 biometrics say (Neckness, Chest width, Body size, Head length and Head width). In addition to Mean Squared Error (MSE), as a measure of the absolute difference errors between the true and estimated biometric values, we report Root Mean Squared Error (RMSE) as a standard error metric. The RMSE of predicted values $\hat{y}$ of a regression's dependent variable

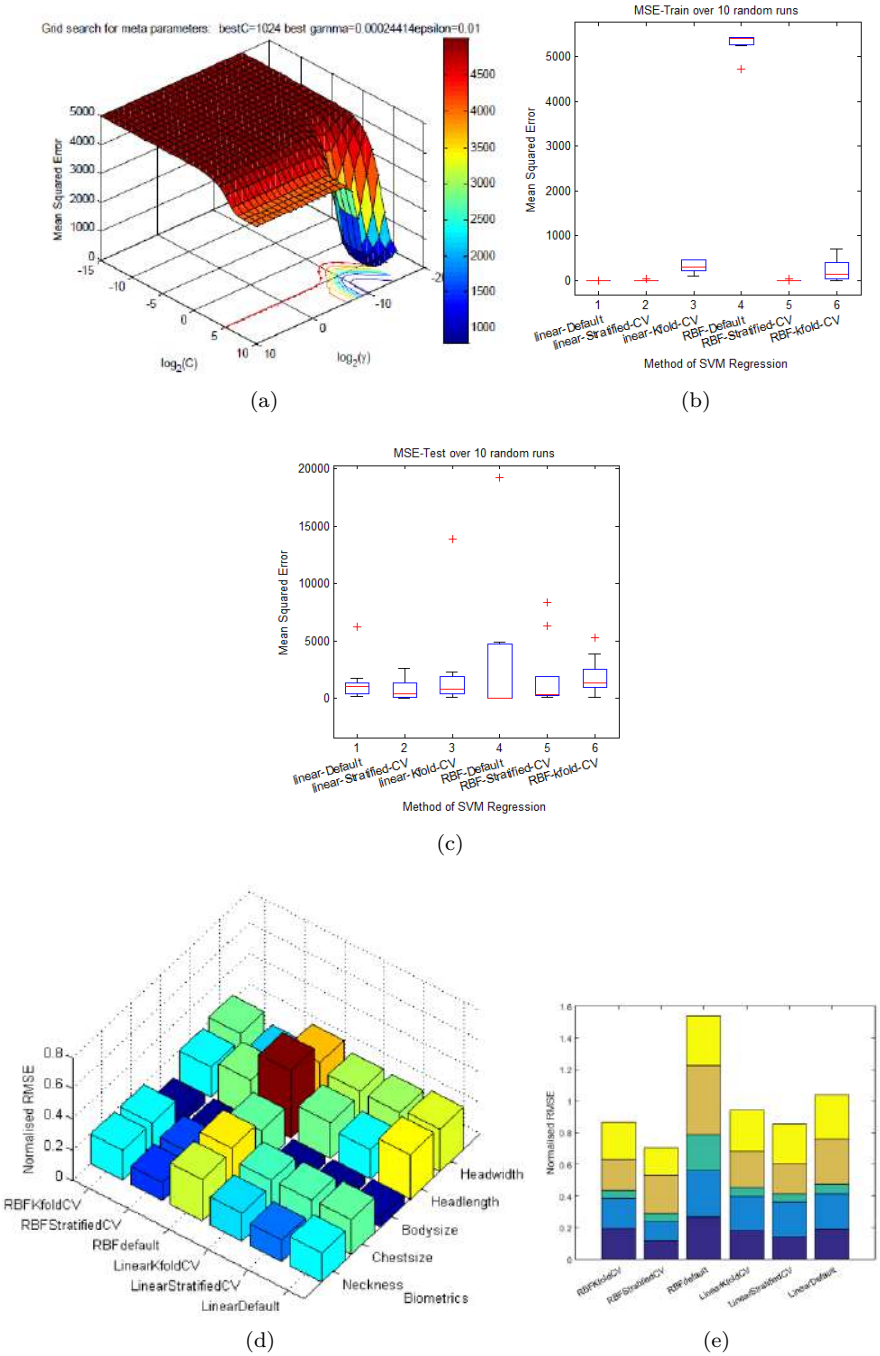


Figure 3.11: (a) Contour and surface plots of the *MSE_Train* distribution for various C and γ meta parameters. Blue corresponds to lowest *MSE_Train*. $log_2(C)$= 10 and $log_2(\gamma)$= -12 produces the least prediction error (lowest *MSE_Train*) (b) MSE for the trainset (*MSE_Train*) over 10 random runs (c) MSE for the testset (*MSE_Test*) over 10 random runs (d) Summary of our various regressors' performance on different biometrics estimation (e) The overall regressor performance on different biometrics estimation.

y is computed for n different predictions as the square root of the mean of the squares of the

deviations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3.13}$$

Normalizing the RMSE facilitates the comparison between datasets or models with different scales. Though there is no consistent means of normalization in the literature, the range of the measured data defined as the maximum value minus the minimum value is a common choice:

$$NRMSE = \frac{RMSE}{y_{\max} - y_{\min}} \tag{3.14}$$

Since the biometrics ranges are different, we leverage *NRMSE* value for the evaluation of our regressor's performance over different biometrics. The visualization of the *NRMSE* values for all the regression methods over all biometrics under consideration, is given in Fig. 3.11(d). Another visualization of the overall regressor performance is also depicted in Fig. 3.11(e).

Table 3.3: Performance of Linear and Nonlinear regression models studied in this work on biometric1 (Neckness), for different parameter settings as well as cross validation schemes. The experimental results over 2-fold as well as 4-fold cross validation settings are shown below. Least values of *MSE_Train* and *MSE_Test* are shown in **bold** characters and the second least in *italics*.

Expt	Kernel	Parameter Setting	Cross validation (CV)	No. of CV fold	MSE_Train	MSE_Test	Meta parameter
(1)	Linear	Default	-	-	***0.0100***	1207.5	default
(2)	Linear	Gridsearch	Stratified	2 4	17.2178 94.1696	***991.1524*** 1319.5	*C= 0.5*
(3)	Linear	Gridsearch	K-fold	2 4	95.24996 468.1726	1347.6 2986.8	*C= 0.25*
(4)	RBF	Default	-	-	4726.4	19211	*default*
(5)	RBF	Gridsearch	Stratified	2 4	0.3974 **0.00010061**	**985.2449** 1240.0	*C=1024; γ=0.00024414*
(6)	RBF	Gridsearch	K-fold	2 4	41.5721 0.5671	1033.1433 1924.0950	*C=1024; γ= 4.8828e-4*

Following are the main findings from our experiments conducted above:

1) *Grid-search on meta parameters can fine tune the measure of fit, and thus the optimal nonlinear experiments outperforms the others in all the cases:* Our experiments with linear function and RBF kernel show that kernelization gives a slight improvement in performance. For e.g., referring to Table 3.3 and Fig. 3.11(b) and (c), we can observe that the default values of parameters produce worse results for RBF kernel (worst results among all) whereas the grid search of the meta parameters could fine tune the performance. Similarly, applying grid search in linear regression also can reduce the estimation error to some extent.

2) *We observed a nearly linear relationship between SC descriptors and the corresponding Biometrics; In other words, Linear Regression lies very close to cross validated nonlinear regression modalities:* In the boxplots of Fig. 3.11 (b) and (c) as well as the barplots in Fig. 3.11(d) and (e), we could observe that linear kernels, as well as the cross-validated RBF kernel produced more or less the same range of estimation errors. Also in terms of consistency of estimation errors also similar results are observed. So we conclude that, the nonlinear kernelization could bring only a small advantage over purely linear regression against our descriptor set. This intuitively indicates that there exists a nearly linear relationship between the Shape Context

descriptor and the corresponding Biometrics.

3) *Cross validation influences in the system performance:* Among two types of cross validation schemes we applied on our data, we could observe that Stratified k-fold CV outperforms the k-fold CV, in terms of accuracy and consistency. After learning the relationship among the Shape Context descriptors and the Soft Biometrics, we built the best regression model for our Re-ID system using *Nonlinear regression with RBF stratified CV*.

Table 3.4: Chart showing the meta parameters settings of RBF for the best regression performance.

Index	**Biometric**	**Kernel type(t)**	**Cost (C)**	**Gamma (γ)**	**Epsilon (ϵ)**
1	Neckness*(N)*	RBF	1024	0.00024414	0.01
2	Chestsize*(C)*	RBF	1024	0.00048828	0.01
3	Bodysize*(B)*	RBF	128	0.00024414	0.01
4	Headlength*(HL)*	RBF	64	0.00048828	0.01
5	Headwidth*(HW)*	RBF	64	0.00048828	0.01

After we learnt the relationship among the Shape Context descriptors and the Soft Biometrics, we built the best regression model for our Re-ID system (see Fig. 3.2) based on the results achieved from our experiments - *Nonlinear regression with RBF stratified CV*. We designed our regression block using the best meta parameters obtained for each biometric in the simulator, and predict the counterpart biometrics in real world dataset using them. The RBF kernel basis meta parameters for each biometrics is provided in Table 3.4.

3.4.3 Re-identification from verbal queries

Refering to system architecture in Fig. 3.2, human query based categorization is related to ***Scenario#2***. Here, the input to the system is a human query specifying the biometric features of the probe, rather than an image query. With this query, our system will not produce a unique *human#ID* as if working with a Re-ID ***Scenario#1***. Instead, the output will be a set of people belonging to that particular category according to the probe description.

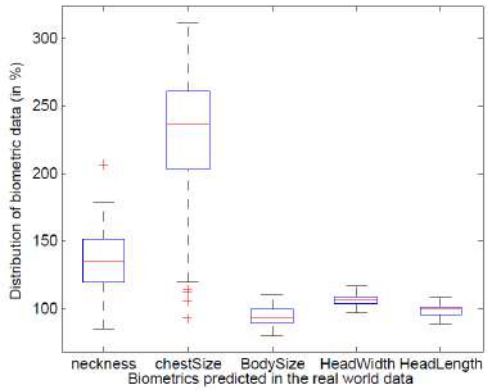


Figure 3.12: Biometric data distribution predicted for the real human population, using the regression model learned using simulated avatars.

As explained earlier, the input is a human query conveying some qualitative information

regarding the biometric features of the person. The regression coefficients obtained from the *Generic* avatar regression model is applied to the SC descriptors of the real human silhouettes in the gallery database and corresponding biometrics are estimated and stored in a gallery database of soft biometrics. Our system collects this gallery dataset of soft biometrics and analyse the distribution of the estimated biometric data in the training population. The most common semantic categories such as *Short (S), Medium (M) and Large (L)* are interpreted in terms of data ranges in this distribution profile. When a human query is available (eyewitness makes a statement regarding the characteristics of the suspect), it is compared against the aforementioned semantic categories, and the valid category of interest retrieved.

Consider a real human dataset as in Figure 3.13. The statistical analysis of estimated biometric values on this dataset is presented in Fig. 3.12. We could observe a range of variances along the biometrics estimated among the dataset. The distribution of Neckness ranges between 90% to 200% of the trained simulator models. Larger necks above that range (like Fig. 3.9(c)) are unexpected in real scenario. The parameter distribution of Chestsize ranges between 100% and 300%, with median close to 220%. This makes sense while checking with similar avatar models in Fig. 3.9(d), which is a common candidate in the real world. Bodysize, Headwidth and Headlength are centered near the 100%, and have lower variances.

It is important to have certain biometrics with large variance in the population in order to avoid the problem of subject interference and to improve the distinctiveness among people. They act as the most discriminative features. One interesting fact to notice is that, from the survey results in Table 3.1, we observed the variance in the chest width, viz., bideltoid breadth[8](2.59) is larger compared to the others. In our sample real world population in Fig. 3.12, we could also observe that chest size shows large variance and happens to be very good discriminative feature.

At the same time, head length and width do not show the same level of variances. A very similar analysis was reported in the real human dataset in Table 3.1, showing smaller variances for head length (0.72) and head width (0.60). These are very interesting observations highlighting the intuitive fact that, our regression model trained in virtual world, could generate similar test result statistics in the real world.

3.4.4 Person Re-ID in real world

We conduct a sample test of person retrieval. Consider a sample real world human dataset of 10 people shown in Fig. 3.13, taken from RGB-D Person re-identification Dataset[9]. We assume 3 categories among the population for each biometric viz. *Short* (***S***-less than lower quartile), *Medium* (***M***-lower quartile to upper quartile) and *Large* (***L***-above upper quartile). For example, in search of a person with large chest size, we try to retrieve the people whose $chestsize \geq 260\%$, which is more than the upper quartile of the distribution. Similarly, for

[8]Bideltoid breadth is same as the chest width we denoted.
[9]http://www.iit.it/en/datasets-and-code/datasets/rgbdid.html

the short chest, we can identify the people category $chestsize \leq 210\%$, which is less than the lower quartile in the data distribution profile. The result thus retrieves a ranked list of people trained in those respective category along with their *#ID*s. Retrieval based on chest query is depicted in Figure 3.14. According to the results, human index #6, #7 and #8 were classified with *Large* chest, and #2, #3 and #5 found to have a *Short* chest. The retrieval based on other parameters is analogous. Albeit, we considered the aforementioned 3 classes (*S,M and L*) as default in our case study, it could be reduced to 2 classes or increased to 4 or more classes (like *XXS, XS, XL, XXL*). The selection of the number of classes and range for each class are the choice of the operator. According to the requirement, he can either split or merge classes. However, with the increase in the number of classes, the retrieval performance decreases due to increase of noise influence in class assignment and inter-class ambiguity.

Among 10 people sample test set each with 4 samples, our retrieval rate for each biometric feature is given in Table 3.5. Since there is no availability of the ground truth for the performance evaluation in the real human dataset, we rely on visual inspection of the probe images and define our *ground truth (GT)*. The rate of correct category retrieval obtained for each person with respect to the ground truth is denoted in retrieval rate. Average retrieval accuracy is found to be the highest for Chest size (77.5%), thus proving to be the best discriminative features among the biometrics.

(a) **#P1** (b) **#P2** (c) **#P3** (d) **#P4** (e) **#P5** (f) **#P6** (g) **#P7** (h) **#P8** (i) **#P9** (j) **#P10**

Figure 3.13: A sample real world dataset for the retrieval test based on human queries on biometric info.

(a) **#P6** (285%) (b) **#P8** (276%) (c) **#P7** (263%) (d) **#P5** (157%) (e) **#P3** (160%) (f) **#P2** (185%)

Figure 3.14: Human categorization based on biometric query: The results for Large chest (L) query and Short chest (S) query are presented in (a)-(c) and (d)-(f) respectively. The retrieved ranked list of human *#ID*s along with the predicted Biometric data value are shown.

3.5 Summary

In this work, we presented a novel proposal towards identifying people in a video surveillance system either through the multimedia data acquired via video cameras or solely by means of manual queries describing natural human compliant labels known as soft biometric traits. Our

Table 3.5: Results of Person retrieval based on Biometric feature vectors estimated by regression. GT refers to the Ground truth biometrics defined by manual inspection, and Retrieval rate is the rate with which our retrieved category agrees with that of Ground truth.

Person index	**Neckness(N)**		**Chestsize(C)**		**Bodysize(B)**		**Headwidth(HW)**		**HeadLength(HL)**	
(*#ID*)	GT	Retrieval rate	GT	Retrieval rate	GT	Retrieval rate	GT	Retrieval rate	GT	Retrieval rate
#1	M	0.25	M	1	M	1	M	0.75	M	0.25
#2	L	1	S	0.5	S	0.5	S	0	L	1
#3	L	0	S	1	M	0.5	M	0.25	M	0.25
#4	M	1	M	1	M	0.5	M	1	M	0.75
#5	L	0	S	0.25	S	1	S	1	M	0.25
#6	S	0.5	L	1	L	0	L	0	S	0.75
#7	L	1	L	1	M	0.75	M	0.5	M	0.25
#8	S	0	L	0.25	L	0	L	0	S	0
#9	S	1	M	1	L	1	M	0	M	0.75
#10	M	1	M	0.75	M	0.5	S	0	L	0
AverageAccuracy	57.5%		77.5%		47.5%		35%		42.5%	

automatic dual mode system is found quite appropriate in the search of an incident happened in a video surveillance, where the security personnel could opt collecting either multimedia info from the camera or eyewitness description of the suspect, which are the common ways of manual person Re-ID.

We introduced a novel feature descriptor, SC descriptor extracted on the head-to-torso region on frontal human silhouettes. Then, the relationship between SC descriptors and soft biometrics, was analysed via Support Vector Regression (SVR), leveraging both linear and nonlinear regression kernels. Such a Regression phase filled the gap between the manual and machine interpretation of human profile and equipped the system to retrieve the person merely by soft biometric description of the subject. In order to provide the best model for the regression analysis, we conducted an extensive study on the impact of various regression schemes, as well as cross validation schemes on SC- BF pairs of our simulated dataset of virtual reality avatars. We observed that the grid search for the best meta parameterized model can fine tune the system for the best performance. In our experiments nonlinear kernel (RBF) basis with stratified cross validation excels in performance compared to all the other schemes. Interestingly, linear regression models are also found to provide good and fast results. This gives us the intuition that the correlation between the SC and biometrics are nearly linear. We substantiated the performance of our system by carrying out person retrieval not only in the simulated platform, but also in a sample real surveillance database. In future work, we plan to extrapolate the feature extraction over full body and to exploit a large set of soft biometrics. Also, we will combine other modalities (e.g., color, texture, face, gait) along with soft biometric features using multimodal fusion techniques.

Chapter 4

Gait based Person Re-ID

High'st Queen of state, Great Juno comes; I know by her gait

— Shakespeare, The tempest

4.1 Gait for person Re-ID

Over the last decade, a number of gait analysis techniques have been proposed towards Person Re-identification/ Recognition. Re-Identification (Re-ID) is associated with change in appearance (carrying bags and different clothings etc.) and uncontrolled conditions (changes in illumination, pose and background). Recognition is a special case of Re-ID, where there is no apparent change in the appearance of the subject and the operator has much control on the conditions (same camera, no change in pose/ background/ illumination etc.).

In order to have a better comprehension of the state-of-the-art techniques, we conducted a substantial overview of different approaches in gait based re-identification conducted in the past, and produced [Nambiar *et al.* , 2016a] by compiling the literature survey findings. The survey paper presents a review of the work done in gait analysis for re-identification in the last decade, looking at the main approaches, challenges and evaluation methodologies.

In classical gait analysis, the most commonly used views are lateral and frontal views. Most of the state-of-the-art techniques address the lateral case, in which the gait can be better observed. Lateral views have the advantage of minimizing perspective distortion and the amount of self occlusion, however, they cannot be applied in narrow passages, since very few gait cycles are observed in these conditions. Hence, in many real world scenarios like indoor narrow corridors and confined spaces, systems that rely on frontal gait analysis are preferred due to the convenience to be installed in confined spaces, as well as the capability to capture longer video sequences, at the same time impose more challenges in terms of perspective and occlusions.

In this thesis, we propose a novel framework for model-free and frontal gait analysis for person Re-ID, by amalgamating the HOF [Dalal *et al.* , 2006] into the the framework of Gait

Energy Image GEI [Han & Bhanu, 2006], and building upon the advantages of both representations. First, the HOF represents the dynamic gait characteristics by encoding the pattern of apparent motion of the subject in a visual scene. Second, GEI enables to average the energy information over a gait cycle to obtain the spatio-temporal gait signature. Our major contributions are as follows:

- A new technique (termed as HOFEI) for person Re-ID in frontal videos leveraging optic flow features.
- Our proposal does not require binary silhouettes, instead computes global dense motion descriptors directly from raw images. This not only bypasses the segmentation and binarization phases, but also facilitates online Re-ID.
- Proposal of a new gait period estimation directly from the temporal evolution of the HOF computed at the lower limbs.
- The demonstration, of the applicability of an optic flow-based method to frontal gait recognition, to the best of our knowledge, is absent in the literature.

The pipeline of our proposed algorithm is shown in Fig. 3.2, which will be detailed in the forthcoming sections.

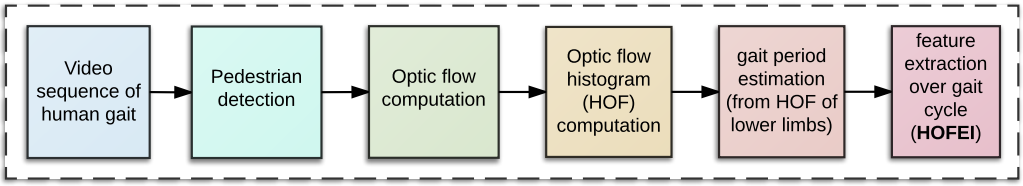


Figure 4.1: Proposed pipeline of the gait analysis.

4.2 Methodology

In this section, the target representation strategy via HOF, gait period estimation and the generation of gait signature Histogram Of Flow Energy Image (HOFEI) are explained in detail.

Histogram of flow: We leverage the Histogram Of Gradients (HOG) encoding scheme mentioned in [Dalal & Triggs, 2005] on the human detection Bounding Box (BB). We provide 2 choices for the human detection BB: either by using the *'Ground truth'* annotations provided, or by using the *'Optic flow'* features to detect the moving section in the image. In this work, we use the default *'Ground truth'* BB. Then, the relative motion distributions of the peripheral human body parts - heads, arms and legs - are described within this BB. In contrast to the original HOG encoding scheme using grid of rectangular cells which overlap, here we use polar cells which better represent the spatial locations of limbs and head along time. Fig. 4.2(a)-(c) show the optic flow computation over a continuous walking sequence of frontal gait and Fig. 4.2(d)-(f) illustrate the sampling scheme of HOF.

When an optic flow image is provided, the first step is to divide it into cells according to the

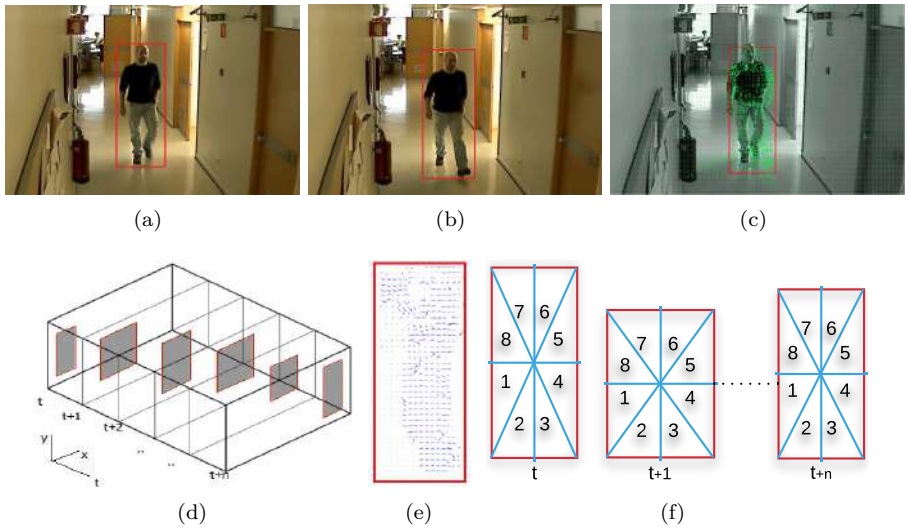


Figure 4.2: Optic flow computation (*top*) and polar sampling scheme for the computation of Histogram Of Flow (HOF) (*bottom*); (a) and (b) show the adjacent video frames of gait. (c) shows the optic flow of the person computed. (d) The cuboid represents a slice of the video sequence spanning a gait cycle (n frames). The shaded regions are the Bounding Box (BB) of the person detected in each frame. (e) A sample of person's optic flow inside the BB. (f) Polar sampling of histogram of flow HOF in each of the images during a gait cycle, whose average results in the HOFEI gait signature.

polar sampling strategy mentioned above, followed by the computation of histogram of flow orientation weighed by its magnitude. Let nR be the number of angular regions (i.e. cells) and nB be the number of bins that define each cell. Hence, the HOF features are parameterised as follows:

$$\mathbf{HOF^t} = \left[HOF^t{}_1 \cdots HOF^t{}_i \cdots HOF^t{}_{nR}\right] \in \mathbb{R}^{nR \times nB} \tag{4.1}$$

where $HOF^t{}_i$ denotes the normalized HOF computed at cell i at frame t. Fig. 4.2 illustrates this, where it is shown a polar sampling scheme with 8 angular cells, $\mathbf{HOF^t}$ is of dimension 64, (with nR=8 and nB=8). We compute the $\mathbf{HOF^t}$ for each frame throughout the video sequence S and the representation for the $\mathbf{HOF_S}$ is expressed as follows:

$$\mathbf{HOF_S} = \left[\mathbf{HOF^t} \mid \cdots \mid \mathbf{HOF^{t+\tau}}\right]^\top \in \mathbb{R}^{\tau \times nR \times nB.} \tag{4.2}$$

where τ denotes the number of frames in the video sequence.

Gait cycle estimation: Humans walk in a periodic fashion. In order to have coherent and reliable gait signature, it is necessary to estimate the gait features over a gait cycle, which acts as the functional unit of gait. A gait cycle is the time period or sequence of events/ movements during locomotion in which one foot contacts the ground to when that same foot again contacts the ground. In our proposal, the estimation of gait period is computed directly from the optic flow measured within the subjects' BB in raw images. This bypasses the computational load related to the traditional image segmentation and other image pre-processing steps in gait period computation.

We extract the periodicity encoded in the HOF sampling cells corresponding to the lower limbs. This choice is motivated by the fact that, the periodic information can reliably be obtained using the dynamic motion cues from the legs. The periodicity of right and left legs induces a similar periodic pattern in its corresponding optic flow. For instance, in a frontal gait sequence, as shown in Fig. 4.3(a) and Fig. 4.3(b), the polar sampling of cells 2 and 3, correspond to the location of legs in the image. More specifically, cell 2 and cell 3 correspond to the right and left leg, respectively.

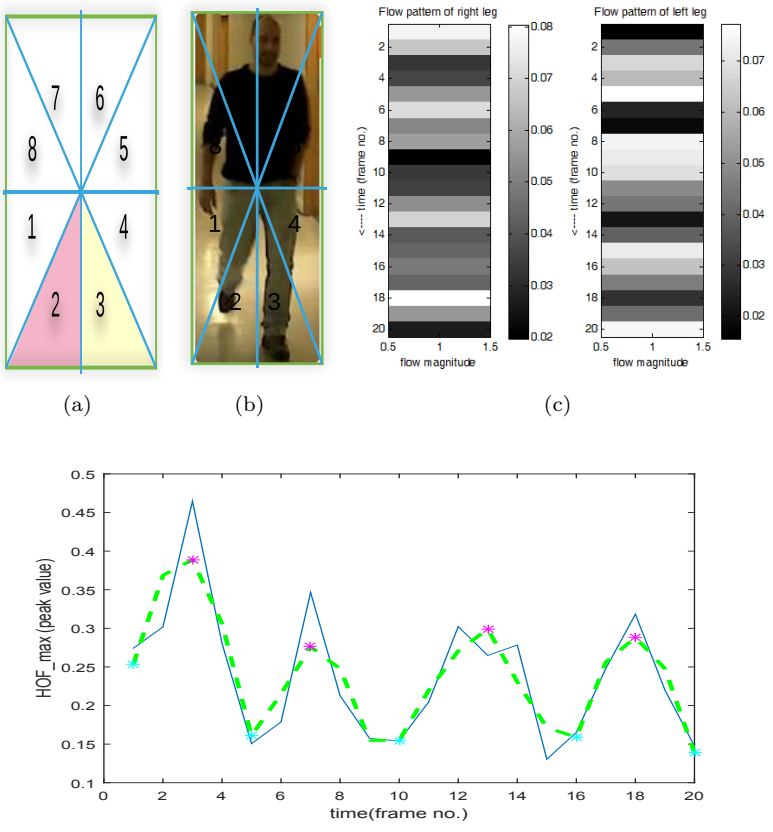


Figure 4.3: (a) Cells 2 and 3 represent sampling cells corresponding to the lower limbs. (b) person's BB under polar sampling scheme depicts that the major area of motion pattern is described by the lower limbs cells. (c) magnitude of the highest peak of the histogram of the right and left legs (cell 2 and 3) during a walking sequence. It is worth mentioning that the minimum value corresponding to the *stance* phase in one leg follows the maximum value corresponding to the *swing* phase in the other leg. (d) estimation of gait period. The frames within two adjacent peaks (in *magenta* markers) denote a gait cycle.

In order to estimate the gait period, we leverage the subset of histogram bins corresponding to cells 2 and 3, i.e., ${HOF^t}_2$ and ${HOF^t}_3$, which represents the lower limbs motion patterns, whose amplitude provides a good signal-to-noise ratio for detection. Then, we compute $\mathbf{HOF^t}$ throughout the video sequence corresponding to either ${HOF^t}_2$ or ${HOF^t}_3$ (since both are complementary). We can notice that this evolution undergoes a periodic pattern as depicted in Fig. 4.3(c),(d). Fig. 4.3(d) shows a periodic sinusoidal curve generated by plotting the HOF peaks of a single leg against the frame (as a function of time). A moving average filter is employed

to smooth the obtained curve measurements (see green dashed curve), and the peaks of the filtered gait waveform allow us to identify the gait cycles. The frames between two consecutive peak points represent a gait cycle. Fig. 4.3(c) visualizes the simultaneous evolution of the HOF pattern peaks of both legs i.e., the amplitude of the highest peak in the histogram of each corresponding leg over time, are complementary since stride phase in one leg is accompanied by the stance phase in the other and vice versa.

Histogram of flow Energy Image: Based on the gait period estimation, as well as the HOF features over video sequences, we compute Histogram Of Flow Energy Image (HOFEI), which is used as the key descriptor of each person. Inspired by the GEI scheme, HOF energy image is obtained by averaging the $\mathbf{HOF^t}$ representations over a full gait cycle, as follows:

$$\mathbf{HOFEI} = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \mathbf{HOF^t} \tag{4.3}$$

where t_1 and t_2 are the beginning and ending frame indices of a gait cycle and $\mathbf{HOF^t}$ is the histogram of flow of the person at time instant t, as defined in Equation. (4.1). More intuitively, the **HOFEI** gait signature provides the relative motion of each body part with respect to the other, over a complete gait cycle.

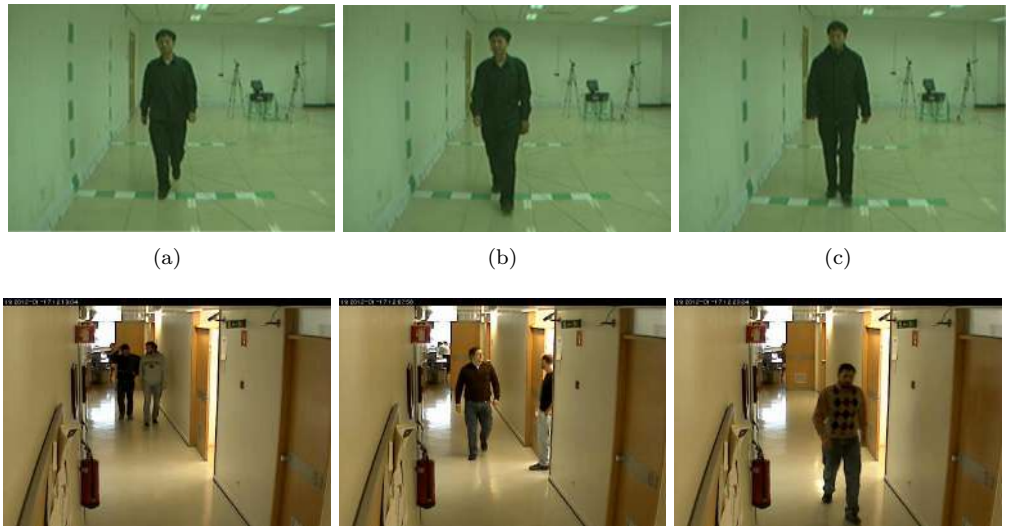
4.3 Experimental Results

Experiments are conducted in two scenarios: Re-ID in controlled scenario vs Re-ID in uncontrolled (busy office) scenario. For the former, we use CASIA dataset B which contains multiple videos of subjects including normal and apparel change (*bag*, *overcoat*) conditions, which makes it suitable for Re-ID scenario. Nevertheless, there is much control over the pose, illumination and background. Hence, it is also suitable to study the recognition of the subject under similar conditions. Hence, we conduct an extensive study on both the re-identification as well as recognition analysis in CASIA dataset. After this feasibility analysis, we apply our algorithm on a more realistic dataset (HDA Person dataset [Nambiar *et al.* , 2014b]) which is used for benchmarking video surveillance algorithms. In contrast to the CASIA dataset, HDA provides uncontrolled environment conditions (change in illumination, pose changes and occlusions), as well as lower frame rate (5fps) similar to a real world video surveillance system, which enables to conduct a Re-ID task in realistic scenario.

4.3.1 Re-ID in controlled scenario : CASIA dataset

CASIA is one of the largest database available for gait recognition and related research [1]. Among the available four different datasets, we used Dataset B for our experiments. Dataset B is a large multiview gait dataset collected indoor with 124 subjects and 13640 samples from 11 different views ranging from 0 to 180 degrees. In our experiments, we consider only the frontal walks (0 degrees), i.e., walking towards the camera. Database B contains three variations, namely view angle, clothing and carrying condition changes, and also presents the human silhouettes for each case. For each person, it contains 10 different video sequences (6 '*normal*'

[1] http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp



(a) (b) (c)

(d) (e) (f)

Figure 4.4: Some sample images from CASIA database and HDA database. (a)-(c) show various appearance (*'normal walk', 'carrying bag', 'wearing coat'*) conditions of subjects in CASIA dataset B. (d)-(f) depict the position of subjects at various distances $\mathbf{D_{far}}$, $\mathbf{D_{middle}}$ and $\mathbf{D_{near}}$ respectively.

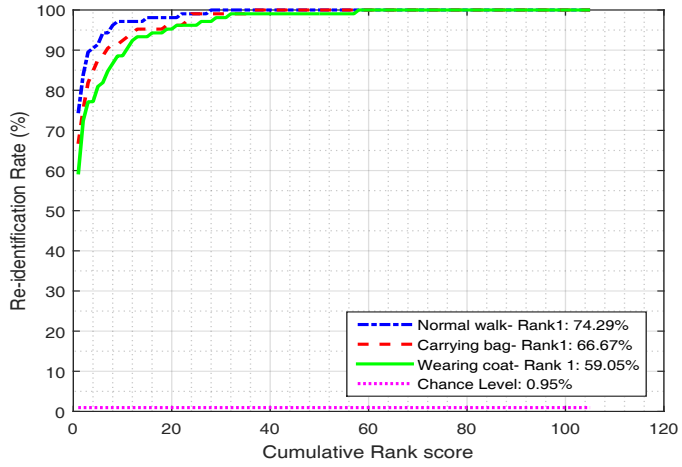
walk, 2 '*bag carrying*' walk and 2 '*overcoat wearing*' walk). Please refer to Fig. 4.4(a)- (c) for samples from CASIA dataset.

In order to evaluate the performance of our system towards long term re-identification, we conduct experiments not only under normal scenario, but also in the apparel change situations such as wearing coat/ carrying bag. For each of these experiments we considered 105 subjects, out of all the available 124 subjects. Videos in which the optical flow information can not be successfully extracted are excluded. For each of these available 1050 videos, we could get at least 3 gait cycles, in order to have enough data for training and testing. Then, for each gait cycle, the corresponding HOFEI is extracted. Regarding the dense optical flow computation, we use Stefan's implementation [2], which provides robust flow estimation by various methods of which, we select the Lucas- Kanade method [Lucas & Kanade, 1981].

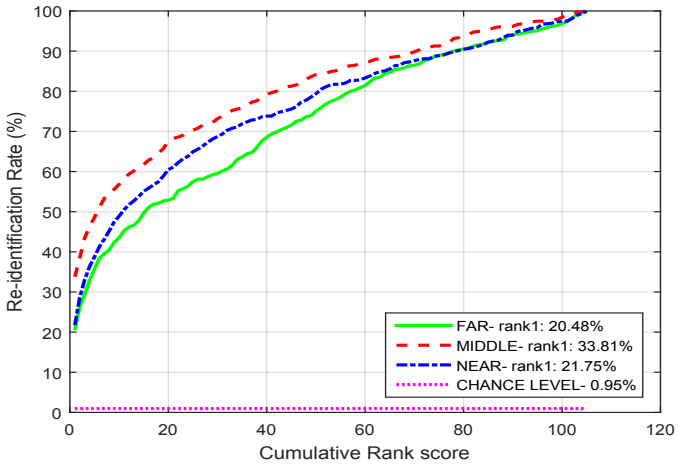
Three main experiments are carried out in this dataset: First is to verify the recognition performance under the same appearance and similar conditions. Second experiment is the Re-ID test conducted in order to verify performance under different appearance conditions. And the third experiment is to test the influence of the distance of the subject in the performance of our system.

Experiment 1) Recognition in regular conditions: In this experiment, we only consider the '*normal*' type videos. The first four sequences are used for training and the last two are placed into the probe set. Then for each person's probe sequences, we compute the minimal Euclidean distance between any of the HOFEI descriptors in the probe and those of each person on the gallery. The minimal distance (most similar) gallery sequence is selected as

[2] http://www.mathworks.com/matlabcentral/fileexchange/44400-tutorial-and-toolbox-on\-real-time-optical-flow



(a)



(b)

Figure 4.5: Re-ID results: (a) presents the CMC curves obtained for Experiment 1 and 2 for different probe cases viz., *normal* case, *bag* carrying case and *coat* wearing case. A chance level of 0.95% is also denoted in *magenta*. The Rank1 recognition achieved for normal, bag and overcoat are 74.29% (78 times the chance level), 66.67% (70 times the chance level), 59.05% (62 times the chance level) respectively. (b) depicts the CMC curves obtained for Experiment 3 at various distance probes viz., $\mathbf{D_{far}}$, $\mathbf{D_{middle}}$ and $\mathbf{D_{near}}$. Middle case outperforms the others. (see online version for colours).

the best matching and sets the identity of the recognized person. The distances to the other persons in the gallery are used to provide a ranked list of identifications, for evaluation. Blue dotted curve in Fig. 4.5(a) shows the Correct Classification Rate (CCR) of this experiment, in terms of Cumulative Matching Characteristic (CMC) Curve. CMC curve shows, how often on average, the correct person ID is included in the best K matches against the training set for each probe. We could observe that a high CCR rate of 74.29% (78 times the chance level), has been achieved under the regular '*normal* walking' conditions.

A similar evaluation strategy, but using silhouette-based approach, had been carried out in [Chen *et al.* , 2009] in all the view angles in CASIA dataset *'normal'* sequences. In order

Table 4.1: Comparative analysis of our method against silhouette-based approaches in [Chen *et al.* , 2009], for the frontal gait sequences of CASIA dataset B. The proposed method (HOFEI) is shown in bold letters.

GMI	GHI	**HOFEI**	GEI	FDEI
68.5%	71.8%	**74.3%**	91.1%	95.2%/100%

to conduct a reasonable comparison with our approach, we select the frontal view results they obtained by using various gait features (GEI, GHI, GMI, FDEI). Table 4.1 shows the comparison results of our strategy (HOFEI) against them in the ascending order. We can observe that CCR of our approach lies in between the others. The higher performance of FDEI and GEI could be attributed to their usage of segmented binary silhouettes and a more powerful classification method (HMM), whereas we use more flexible optic flow based features and a simpler classification method (Euclidean distance). Therefore, we consider the proposed feature competitive with the state-of-the-art, while more versatile. Since it does not require any pre-segmentation phase, it is easier to use in automatic RE-ID systems.

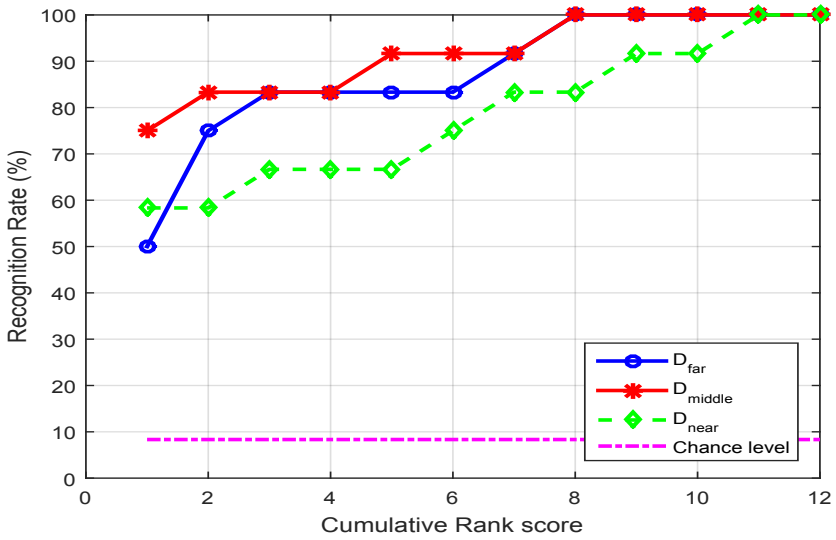
Experiment 2) Re-identification under change in appearance: In this experiment, we use all the 6 *'normal'* type videos for training, and *'wearing coat'/ 'carrying bag'* type videos for testing. In the '*bag*' case, we keep both the bag carrying sequences as the probe whereas all the 6 *'normal'* video sequences as the training set. A similar method is employed for *overcoat* scenario as well. Classification is similar to Experiment 1 (NN classifier+ Euclidean distance). The recognition results obtained are presented in Fig. 4.5(a). The apparel change recognition rates for bag (*red curve*) and coat (*green curve*) scenarios are 66.67% (70 times the chance level) and 59.05% (62 times the chance level), respectively. The lower CCR of *overcoat* condition could be ascribable to the global change in the flow features, whereas the *bag* either influence only a local flow change, or being occluded in some cases (occluded by hand, as in Fig. 4.4(b) or occluded while wearing as a backpack). No similar results in the appearance change conditions have been encountered in [Chen *et al.* , 2009] for comparative evaluation.

Experiment 3) Variable distance to camera: Here, we are testing the robustness of the system when the subject is at different distance to the camera. In frontal sequences, the variability of the gait features with distance may have a significant impact on performance. Here we study the ability of the method in recognizing persons at a distance for which there are no gallery examples. We consider the '*normal'* type of videos for this experiment. In order to verify the impact of different distances, we conduct 3 case studies. In contrast to the previous experiments carried out on sets of videos, here we are conducting the analysis on each gait cycle instance. Hence, performance will be lower than in the previous experiments, that used all gait cycles in the sequence for the classification. However, in this experiment we are not comparing absolute performance, but relative performance according to camera distance.

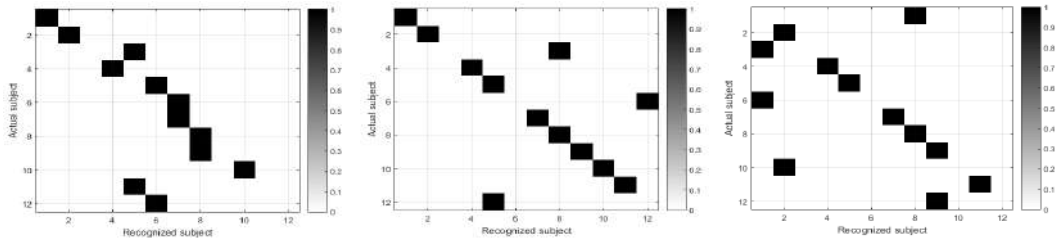
There are minimum of 3 gait cycles in each video sequence. In the first case study we keep all the *'normal'* gait cycle snippets seen at far distance $\mathbf{D_{far}}$ as the probe. The training set in this case is the *'normal'* $\mathbf{D_{middle}}$ and *'normal'* $\mathbf{D_{near}}$. Hence per person, we have 6 $\mathbf{D_{far}}$ probe and 12 training set ($\mathbf{D_{middle}}$ and $\mathbf{D_{near}}$). Then, in the second case study, the $\mathbf{D_{middle}}$

is considered as the probe and $\mathbf{D_{far}}$ and $\mathbf{D_{near}}$ are kept as the training sets. Similarly, in the third case study, $\mathbf{D_{near}}$ videos are the probe and the others are kept as the training set. The Re-ID results are shown in Fig. 4.5(b). We can observe an expected drop in the CCR rate while conducting Re-ID with each gait cycle as the probe in this Experiment 3, rather than sets of videos as the probe in Experiment 1 & 2. $\mathbf{D_{middle}}$ case outperforms the other two cases, with 33.81% rate (35 times the chance level) whereas the far and near cases have recognition rates 20.48% (21 times the chance level) and 21.75% (22 times the chance level) respectively. In the case of $\mathbf{D_{middle}}$ as the probe, higher recognition rate could be attributed to the fact that, trained on the extreme ranges the classifier performs an interpolation when predicting values for the middle range, whereas in the other two $\mathbf{D_{far}}$ and $\mathbf{D_{near}}$ cases it has to extrapolate to one of the extremes, which is often an ill-posed operation.

4.3.2 Re-ID in uncontrolled scenario: HDA Person Dataset



(a)



(b) (c) (d)

Figure 4.6: Recognition results: (a) presents the CMC curves obtained on 3 different probe cases viz., $\mathbf{D_{far}}$, $\mathbf{D_{middle}}$ and $\mathbf{D_{near}}$. A chance level of 8.333% is also denoted in *magenta*. The Rank1 recognition achieved for $\mathbf{D_{far}}$, $\mathbf{D_{middle}}$ and $\mathbf{D_{near}}$ are 50% (6 times the chance level), 75% (9 times the chance level), 58.33% (7 times the chance level) respectively. (b)-(d) show the confusion matrices for the 3 probe cases $\mathbf{D_{far}}$, $\mathbf{D_{middle}}$ and $\mathbf{D_{near}}$ respectively.

HDA dataset [Nambiar *et al.* , 2014b] [3], is a labelled image sequence dataset for research on

[3] http://vislab.isr.ist.utl.pt/hda-dataset/

high-definition surveillance. The dataset was acquired from 13 indoor cameras distributed over three floors of one building, recording simultaneously for 30 minutes during a busy noon hour inside a University building. Among the 13, we select only a single camera recording (Camera19), containing frontal gait sequences. The camera has the VGA resolution of 640×480, with a frame rate of $5fps$. In this experiment we considered 12 people that crossed the whole corridor, and for which we could get at least 3 gait cycles in order to have enough data for training and testing. We collect each subject's walking frames, and from them we extract minimum three gait cycles and their corresponding **HOFEI**. Unlike the CASIA dataset, HDA is uncontrolled scenario since it contains varying illumination conditions during the walk, changing backgrounds, break points in between the walks (entry/ exit in the room along the way), occlusions by other person/ wall/ image boundary, self occlusions, slight changes in the pose and limb movements during the walks.

Due to the limitation of larger video sequences as well as varying appearance conditions per person, we exclude the CASIA counterpart Experiment 1 and Experiment 2 in HDA dataset. Here we only conduct Experiment 3, quite similar to the one carried out in CASIA dataset. We consider three cases in which we compute the **HOFEI** descriptor: far ($\mathbf{D_{far}}$), middle ($\mathbf{D_{middle}}$) and near ($\mathbf{D_{near}}$) sequences, as depicted in Fig. 4.4(d)-(f). Under this set of descriptors, we perform a leave-one-out evaluation where one set is kept as the probe and the other two sets as the gallery (i.e, a total of three trials). Thus, in each trial we have 24 training descriptors in the gallery and we test against 12 test probes. Then, each test sample will search for the minimal Euclidean distance between itself and the gallery descriptors, under the nearest neighbor classification method. Fig. 4.6 demonstrates the recognition results in terms of Cumulative Matching Characteristic (CMC) curve and confusion matrix. The highest Rank-1 recognition rate of 75% (9 times the chance level) is achieved while using $\mathbf{D_{middle}}$ as the testing data. At the same time, the Rank-1 accuracy achieved by the test sets $\mathbf{D_{far}}$ and $\mathbf{D_{near}}$ are 50% and 58.33% respectively.

Referring to the CMC curve, another interesting observation is that the cumulative recognition rate improves drastically for both $\mathbf{D_{middle}}$ as well as $\mathbf{D_{far}}$ cases in comparison with $\mathbf{D_{near}}$, with the number of trials. This accentuates that gait sequences are better observed in far sequences than the closer ones since video frames close to the camera may undergo occlusions and thus result in poor encoding of the body flow features.

4.4 Summary

We analysed the potential of exploiting histogram of optic flow for frontal human gait analysis for person Re-Identification (Re-ID). The main advantage of such a methodology is that no silhouette segmentation is required and thus can be facilitated towards online Re-ID system. A novel idea of flow based gait period estimation as well as a novel Histogram of Optic flow Energy Image (**HOFEI**) over the entire body are proposed in this work. We experimented the proposed framework upon a controlled benchmarking gait dataset (CASIA dataset) and a more unconstrained, thus harder, benchmarking video surveillance dataset (HDA Person dataset). We verified the effectiveness of the proposed method in both cases, under very

different background clutter and sampling rates (25Hz in CASIA vs 5Hz in HDA). Extensive studies were conducted in CASIA dataset, i.e., regular case, change in appearance and influence of variable distance. Promising results were reported in each experiment, showing a Re-ID rate of 74.29% (78 times the chance level) in the *normal* scenario. In HDA dataset person Re-ID also a good performance rate of 75% (9 times the chance level) was reported, under different camera distance conditions. In future work, we plan to extrapolate this work towards pose invariant person re-identification scenario.

Chapter 5

Towards view-point invariant Person Re-ID

In theory there is no difference between theory and practice. In practice there is.

— Yogi Berra

5.1 Introduction

As we have already mentioned in the dissertation, Re-ID has many challenging issues that result from the high variability of the people's appearance in the camera images due to different illumination, different clothes, occlusions, postures and camera's opto-electric characteristics and perspective effects. Among them, pose is the most challenging one since it creates drastic changes in the feature in different directions. In order to tackle such scenarios, certain pose invariant Re-ID scheme has to be learned.

With this in mind, we propose towards pose invariant gait based Re-ID leveraging 3D information. In order to facilitate this, we employ existing cutting edge technologies such as KINECT systems to collect quite precise and detailed information of the human dynamics in the scene. The key advantage of use KINECT like systems is that it can collect 3D skeleton data that are view-invariant and scale-independent, along with other sensory information viz., color and depth. So, in this work, we employ 3D KINECT data to facilitate towards pose invariant long term person Re-ID.

Another key idea to incorporate is multi-modal fusion i.e., to fuse either multiple biometric features or biometric+ appearance features, in order to enhance the Re-ID system performance. Some traditional fusion strategies towards multi-modal fusion has been explained in the handbook of Multibiometrics by [Ross, 2007]. According to that, there are many levels at which the fusion could be carried out, i.e., sensor level fusion, feature level fusion, score-level fusion, rank level fusion or decision level fusion. In this work, we also exploit this idea of 'Multi-modal fusion' by leveraging score-level fusion strategy upon different biometric features.

In detail, we propose a biometric enabled person re-identification system, using two kinds

of soft biometric features i.e. anthropometric features and gait features, extracted from the human body skeleton computed by a Microsoft Kinect^{TM} sensor v.2. Anthropometry involves the systematic measurement of the physical properties of the human body, primarily dimensional descriptors of body size and shape. Human gait includes both the body posture and dynamics while walking [Lee & Grimson, 2002]. The cues are extracted from range data which are computed using an RGBD camera. Hence, the great constraint of appearance constancy hypothesis can be relaxed and facilitated towards long-term person Re-ID. To the best of our knowledge only a very limited number of works have been employed in this regard, furthermore, they employ view-point dependent approaches i.e. data is collected and algorithms are tested with a single walking direction with respect to the camera.Barbosa *et al.* [2012b], [Gianaria *et al.* , 2014] and [Andersson & Araujo, 2015]. In this paper, we propose a view-point invariant person re-identification method tested with subjects walking in different directions, by using multi-modal feature fusion of anthropometric and gait features.

The major contributions of the paper are two fold:

- First, to validate the effect of various anthropometric and gait features in distinguishing a person among the population and facilitate towards person Re-ID from those soft-biometric cues. In order to better understand this, we conduct a thorough study by exploiting individual features or combination of features (via fusion).

- Second, is the actual demonstration of the real impact of view-point on the Re-ID paradigm. Since skeleton coordinates provided by kinect data are, in principle, view-point invariant (can be normalized to a canonical view-point by a roto-translation transformation), many works assume view point invariance from the start and do not validate experimentally this assumption. Despite skeleton coordinates are naturally view point invariant, their computation is not (the skeleton reconstruction process depends on view points and self-occlusions). Most work in the literature do single-view probe and single (same)-view gallery (which is basically the view-point dependent approach), which does not allow assessing the view-point invariant characteristics of the algorithm. In order to perform a benchmark assessment, we experiment in this work explicitly different view-points in the probe and gallery samples. In addition, we conduct several tests of view-point invariance: (i) single-view-point probe with multi-view-point gallery (*pseudo* view-point invariance); (ii) novel-view-point probe with multi-view-point gallery (*quasi* view-point invariance) and (iii) novel-view-point probe with single-view-point gallery (*full* view-point invariance). The former two require a large effort in the gallery creation. The latter, is the easiest and most flexible form since only a single camera is required and the person enrollment stage is very simple (one pass only).

This chapter is organized as follows. In Section 5.2, we explain the proposed methodology. In particular, we present the data acquisition set up, feature extraction, signature matching and evaluation methodology. In Section 5.3, we detail the various experiments conducted and the results achieved. We summarize our work and enumerate some future work plans in Section 5.4.

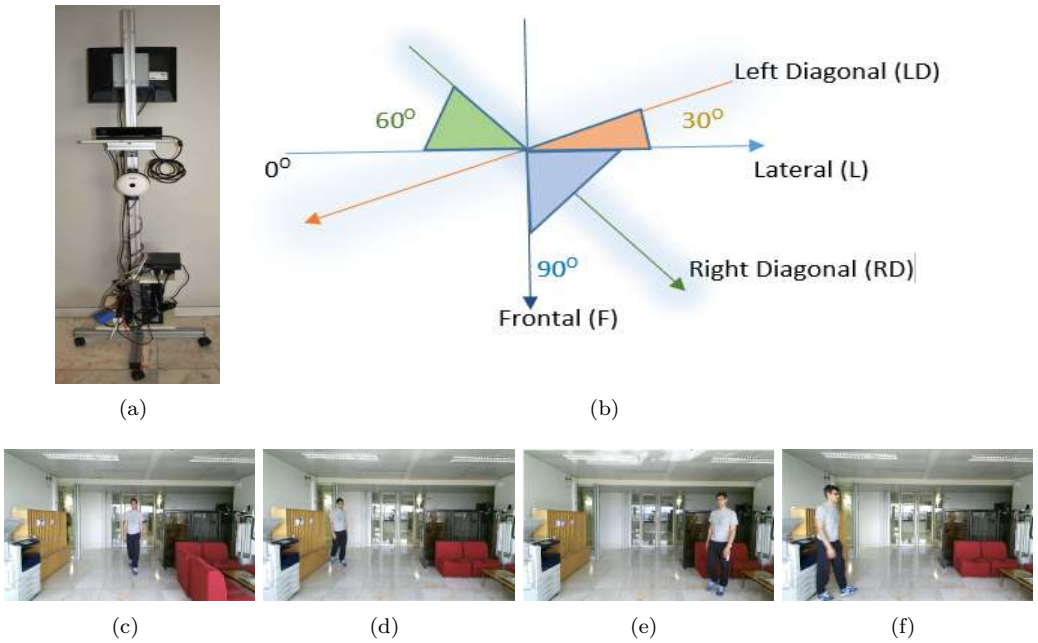


Figure 5.1: Data acquisition: (a) System set up (b) Subject walking directions in front of the acquisition system (c) Sample frames from our data acquisition, in four different directions-frontal(∼90°), right diagonal(∼60°), left diagonal(∼30°) and lateral(∼0°) respectively.

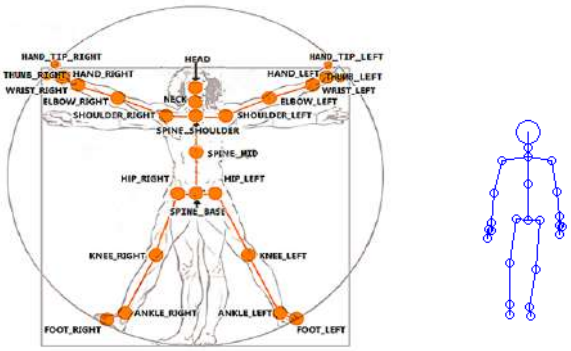
5.2 Methodology

In this section, we explain the data acquisition and proposed methodology. More specifically, we detail the set up and the data collection procedure conducted in the host laboratory. Then, we describe various stages of data analysis including pre-processing, feature extraction, signature matching and experimental evaluation strategies.

5.2.1 Data acquisition set up

For the data acquisition, we used a mobile platform, in which the kinect sensor was fixed at a height of an average human (See Fig. 5.1(a) for the data acquisition system). This mimics normal surveillance scenarios as well as changes in the position of camera over time, as in a long term person Re-ID scenario. The kinect device is composed of a set of sensors, which is accompanied with a Software Development Kit (SDK), that is able to track movements from users by using a skeleton mapping algorithm, and is able to provide the 3D information related to the movements of body joints. We acquired all the three available data i.e. skeleton, colour and depth. Since the proposed gait algorithm employs the skeleton information, it necessitates to be of multiple frames with high frame rate, and hence captured at the full frame rate of the sensor @ 30fps. In this second version of the device, it is able to track 25 joints at 30 frames per second. Colour and depth information are employed for appearance based features, which generally require single frame, and hence was captured at 1fps. However, these were not used in the current work.

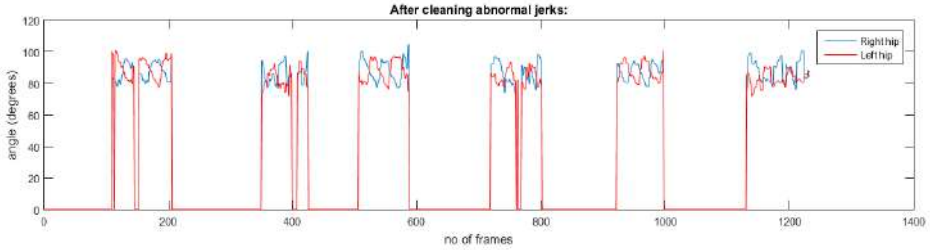
[2]For body joint types and enumeration, refer to the link: https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx



(a) (b)

Figure 5.2: (a) Skeleton positions relative to the human body[2](b) A sample skeleton body visualization from our collection.

(a)



(b)

Figure 5.3: (a) The abnormal shifts towards the ending of each sequence are due to the jerks of skeleton occurring at its respective frames. (b) Abnormal frames are filtered out. Now we have the cleaned frames selected.

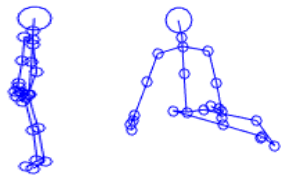
In order to ensure view-point invariance in our acquisition set up, we collected multiple views of 20 subjects in four different directions, along both ways, as shown in Fig. 5.1(b). We define the direction angle with respect to the image plane. Lateral (L) is at ∼0° and Frontal (F) is at ∼90°. And there are two diagonal walks at different view angles. Right Diagonal (RD) begins at one of the corners of the hall, which has ∼60°, whereas Left Diagonal (LD) begins somewhere in the half way, thus defining ∼30°. In each of these four directions, a minimum of three walking sequences were collected both in the front and rear views (refer Fig. 5.1(c)-(f)). During the walking, the people are assumed to walk with their natural gait. Altogether we have 240 video sequences comprising 20 subjects (12 video sequences per person) in the aforementioned directions. Since kinect gets the joint information of the skeleton data, it is in

principle, view-point and scale invariant. In addition to that, we hypothesize that the subject makes straight walks during a single gait acquisitions, as kinect depth range is limited (80cm to 4 meters) .

Kinect can track in real-time a skeleton model, composed of 25 body joints, as shown in Fig. 5.2(a). The skeleton joints can be used to describe the body measurements (anthropometrics) as well as the body movements (gait) in real time and in 3D space [Shotton *et al.* , 2013].

5.2.2 Pre processing

Prior to the feature extraction, we applied some pre-processing for noise removal. The primary effect of noise are jerks/ abnormalities in the skeleton data, during the sequences (see examples in Fig. 5.4). In addition, in some frames, the skeleton is not detected. We could observe that, when the person approaches the boundary of the kinect range, these issues occur very often. In order to handle such situations, we propose a semi- automatic approach to select the best frames to retain and further analyse out of a video sequence.



(a) (b)

Figure 5.4: (a) Some views confuse the joint positions making the skeleton based approach quite difficult (b) Abnormal jerks occuring at certain frames, during the video sequence.

Table 5.1: List of anthropometric and gait features used in our experiments. (L& R correspond to 'left and right' and x& y correspond to 'along x and y axes')

Anthropometric features	**Gait features**	
Height	Hip angle(L&R)	Hip position(L&R)(x& y)
Arm length	Knee angle(L& R)	Knee position(L&R)(x& y)
Upper torso	Foot distance	Ankle position(L&R)(x& y)
Lower torso	Knee distance	Hand position(L&R)(x& y)
Upper-lower ratio	Hand distance	Shoulder position(L&R)(x& y)
Chestsize	Elbow distance	Stride
Hipsize	Head position(x& y)	Stride length
	Spine position(x& y)	Speed

Humans walk in a periodic fashion. It is necessary to estimate the gait feature over each of these periods of walking, known as gait cycle, which acts as the functional unit of gait. A gait cycle comprises of sequence of events/ movements during locomotion since one foot contacts the ground until the same foot again contacts the ground. Prior to getting the gait period, we intend to filter out the unwanted jerks by means of exploiting the evolution of hip angles over time. We noticed that the jerks made these angles to grow abnormally, which also created drastic variations in the corresponding signals. An example of such a situation is depicted in Fig. 5.3(a). In order to clean/ remove such unwanted frames, we put a threshold on the

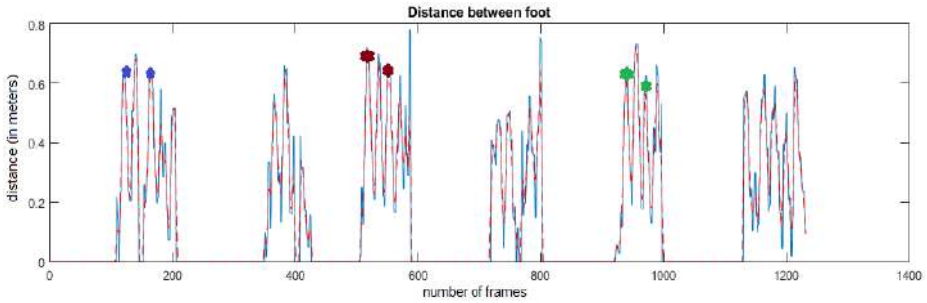


Figure 5.5: Gait cycle estimation. The two adjacent markers (3 consecutive peak) within a sequence, represent a gait cycle.

angular values (usually, the normal expected values of hip angles are in between 70° <hip angle<105°). Only the frames containing the angles in between the upper and lower threshold are selected. This step automatically cleans our noisy data. A cleaned version of the previous signal is depicted in Fig. 5.3(b).

The next step is gait cycle estimation. In order to have a better overview of how the lower limbs move along the video sequences, we compute the distance between the feet during a gait sequence. The three consecutive peaks in such a signal provides a gait cycle. Referring to Fig. 5.5, we can see that in each video sequence, the frames between adjacent markers (stars in same colour) make a gait cycle[3]. At this point, we make this step manually. Albeit we provide the method to automatically select the adjacent peaks defining a gait cycle, we carry out a manual verification by checking the real video sequence and the signal peaks to verify that they are aligned. Also, the phase is verified at this point by checking which leg is in movement. From the peak signal alone, this information is not easy to extract.

After selecting the frames defining gait cycle, we extract the features.

5.2.3 Feature extraction

After data acquisition and filtering, attributes were extracted for each walk, both static physical features defining the anthropometric measurements and dynamic gait features defining the kinematics in walking. To each subject, an identifier was provided for re-identification. The extracted feature attributes are explained in detail, next.

Anthropometric features: Under the anthropometric feature set, we collected many body measurements defining the holistic body proportions of the subject. This includes height, arm length, upper torso length, lower torso length, upper to lower ratio, chest size, hip size. These seven features constitute the body features.

The length of a body part is defined as the sum of the lengths of the links between the delimiting joints. For example, the arm length is the sum of Euclidean distances from shoulder to elbow (joint 4-joint 5), elbow to wrist (joint 5- joint 6) and wrist to hand (joint 6- joint 7). We calculate these static features across each frame, and then compute the mean value of each

[3]Note that, we collect three sequences of walking per person in each direction. Since the person makes a walk in a direction, and then a return walk to the initial point, apparently we have 6 sequences, as we can see in Fig. 5.5. However, we do not consider the return walks in this work, and hence, we have altogether 3 video sequences under consideration, as marked.

feature over a gait cycle. The mean value of the anthropometrics over gait periods, are used as the static feature descriptors in our experiments.

Gait features: Under the gait features, we collect behavioural features, deriving from the continuous monitoring of joints during the gait. The key advantage of using the kinect is to collect a rich set of view-point invariant[4] dynamic spatio-temporal features derived from the body movements.

First we computed three scalar features related to walking, viz., stride length, stride time and the speed of walking. The stride length is the distance between two stationary positions of the same foot while walking (Equation (5.1)). It comprises the left step length and right step length[5]. The duration to complete a stride is called stride time (Equation (5.2)). It is obtained by calculating the number of video frames in a gait cycle divided by the frame rate of acquisition (30 fps). From these two, we can obtain the speed of walking as the ratio between stride length and stride time (Equation (5.3)).

$$\text{Stride length} = \text{Left Step length} + \text{Right Step length} \tag{5.1}$$

$$\text{Stride time} = \frac{\text{Number of frames in gaitcycle}}{30} \tag{5.2}$$

$$\text{Speed} = \frac{\text{Stride length}}{\text{Stride time}} \tag{5.3}$$

In addition, we also computed a set of 32 features, related to the temporal evolution of the angles (at various body joints), distance (between various right-left limbs during the gait) as well as the position (evolution of body joint along the gait). From these spatio-temporal gait signals, we extract the mean and variance of the signal. Altogether, we have a feature set containing 35 gait features (3 scalar and 32 dynamic) and 7 anthropometric features. Table 5.1 presents a detailed list of the feature set.

5.2.4 Signature matching

This section explains how the features can be employed either individually or jointly towards the Re-ID problem. A classical Re-ID problem is usually evaluated by considering two sets of signatures (feature descriptors) collected from people: a gallery set and a probe set. Then, the Re-ID evaluation is carried out via associating each of the signature of the probe set to a corresponding signature in the gallery set.

To evaluate the performance of Re-ID algorithms in closed-set scenarios, the Cumulative Matching Characteristic (CMC) curve [Grother & Phillips, 2004] is the most acclaimed and popular method of choice. The CMC curve shows how often, on average, the correct person ID is included in the best K matches against the training set, for each test image. In other words, it represents the expectation of finding the correct match in the top K matches.

[4] As mentioned before, despite the joint coordinates can be easily transformed to a canonical reference frame, the process to estimate the joints positions suffers from self-occlusions due to view-point.

[5] Step length is the distance between the heel contact point of one foot and that of the other foot.

Nearest Neighbor (NN) is among the most popular as well as most performing classifier, which is commonly used in similar full body biometrics realm [Andersson & Araujo, 2015], [Barbosa *et al.* , 2012b]. Hence, in this work, we exploit NN approach for the classification, using the Euclidean distance as metric. Suppose, we have signatures representing each individual feature vectors, the Euclidean distance between the signature in the probe is compared against the rest in the gallery. Then, the most similar signature in the gallery is selected as the correct Re-ID class.

Concerning anthropometric features in our work, the feature vector is composed of multiple body features, where each of the features has a numerical value associated with an individual trait e.g. height, arm length. In the case of gait features, these individual features are vectors representing mean and variance. Hence, while computing the Euclidean distance, we calculate the distance for each individual feature in the probe, against their corresponding feature peers in the gallery. Thus, we get the Euclidean distance of each probe feature against the gallery, as a distance matrix.

Let us define a probe descriptor $\mathbf{P}$, which is a concatenation of n individual features.

$$\mathbf{P} = [p_1, p_2, \cdots, p_i, \cdots p_n] \in \mathbb{R}^{1\times n} \tag{5.4}$$

The gallery contains a set of similar feature descriptors, which we represent as a matrix G. Each row of G represents an n-dimensional feature vector corresponding to an individual. Likewise, k feature descriptors from multiple subjects are arranged to make a gallery matrix of dimension $k \times n$, as follows.

$$\mathbf{G} = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,i} & \cdots & g_{1,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{j,1} & g_{j,2} & \cdots & g_{j,i} & \cdots & g_{j,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ g_{k,1} & g_{k,2} & \cdots & g_{k,i} & \cdots & g_{k,n} \end{bmatrix} \in \mathbb{R}^{k\times n} \tag{5.5}$$

Then, for the Euclidean distance computation, we calculate the distance of each individual probe feature element, say, $p_i, (i = 1, ..., n)$ against its counterpart feature samples in gallery i.e. $g_{j,i}, (j = 1, ..., k)$, as a distance vector viz., $D(p_i, g_{j,i})$.

$$\begin{aligned} D(p_i, g_{j,i}) &= |p_i - g_{j,i}| \,, \\ \forall\, i &= 1, .., n \;\&\; j = 1, .., k. \end{aligned} \tag{5.6}$$

This results in a distance matrix $\mathbf{D}\in \mathbb{R}^{k\times n}$, as follows in Equation 5.7. Each element in the matrix D is given by $d_{j,i} = D(p_i, g_{j,i})$.

$$\begin{aligned} \mathbf{D} &= \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,i} & \cdots & d_{1,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{j,1} & d_{j,2} & \cdots & d_{j,i} & \cdots & d_{j,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{k,1} & d_{k,2} & \cdots & d_{k,i} & \cdots & d_{k,n} \end{bmatrix} \in \mathbb{R}^{k\times n} \\ &= \begin{bmatrix} \mathbf{d_1} & \mathbf{d_2} & \cdots & \mathbf{d_i} & \cdots & \mathbf{d_n} \end{bmatrix} \in \mathbb{R}^{k\times n} \end{aligned} \tag{5.7}$$

Our idea is to get a single distance score, corresponding to the overall feature set. We accomplish this via a score level fusion strategy. Since different features have different magnitude

ranges, the distance scores also will have its impact. Hence, while doing the fusion, the score will be biased towards the higher measured distance, leading to the problem of heterogeneity of measures. In order to avoid this, we carry out a min-max normalization strategy, which normalize each of the feature distance score within the [0,1] range. More specifically, we normalise each column corresponding to a particular feature, separately, i.e. considering the distance vector corresponding to a particular feature as in Equation 5.7, $\mathbf{d_i} = [d_{1,i}, \cdots, d_{j,i}, \cdots, d_{k,i}]^T$, the normalized distance vector $\mathbf{z_i} = [z_{1,i}, \cdots, z_{j,i}, \cdots, z_{k,i}]^T$ is computed as follows:

$$\mathbf{z_i} = \frac{\mathbf{d_i} - \min(\mathbf{d_i})}{\max(\mathbf{d_i}) - \min(\mathbf{d_i})} \tag{5.8}$$

Afterwards, we generate the fused feature score $\mathbf{Z}$, by summing the individual normalised distance vectors, $\mathbf{z_i}$ with $i = 1, ..., n$.

$$\mathbf{Z} = \Big[\mathbf{z_1} + \mathbf{z_2} + \cdots + \mathbf{z_i} + \cdots + \mathbf{z_n}\Big] \in \mathbb{R}^{k \times 1} \tag{5.9}$$

Then, we sort the fused score $\mathbf{Z}$ in the ascending order and calculate the final CMC curve based on the ranked list of matches.

5.2.5 Evaluation methodology

In order to evaluate our proposal, we conduct multiple extensive experiments to verify the impact of each feature individually and jointly, as well as the influence of various view-points on the Re-ID paradigm. Basically, we conduct two major experiments in this regard. 1) view-point dependent and 2) view-point independent.

In the view-point dependent Re-ID experiment, the walking direction is pre-defined. Hence, the gallery and probe contains the samples from the subjects with the same walking direction. Apparently, this is a much simpler problem of person recognition[6]. In this view-point dependent experiment, further detailed analysis is carried out in order to understand the impact of various features (individual vs fusion) on the overall Re-ID.

In the view-point independent Re-ID experiment, the key idea is to corroborate the effect of different walking directions in the Re-ID scenario. We categorize three major view-point invariant scenarios in this regard -a) Pseudo view-point invariance, b) Quasi view-point invariance and c) Full view-point invariance- based on the samples available in the gallery and probe sets (See Table 5.2). The Re-ID becomes more challenging while moving from pseudo towards full view-point invariant, due to the limited availability of samples in the training set as well as the challenging view angles in the probe set.

5.3 Experimental Results

Since a standard gait dataset with different views acquired with kinect sensor was unavailable, we created a new one consisting of 20 people walking in four different directions i.e. Frontal (F),

[6]Recognition is a special case of Re-ID, in which the operator has much control on the conditions (same camera, no change in view-point/ illumination/ background etc.)

Table 5.2: Chart showing the Re-ID accuracy rates for Experiment 4.2.2

Index	**View-point invariance**	**Gallery**	**Probe**
a	Pseudo	Multi views	Single view
b	Quasi	Multi views	Novel view
c	Full	Single view	Novel view

Left Diagonal (LD), Right Diagonal (RD) and Lateral (L). We have asked each person to walk naturally along a hall in four directions, and three times in each direction. Thus, altogether we have 12 sequences per person in different directions i.e. a total of 240 sequences in the dataset.

In this work, we conduct multiple experiments, as explained in Section 5.2.5. In the first experiment, we conduct Re-ID in individual directions, and in the second experiment, we employ view-point invariant Re-ID. In each of these experiments, we evaluate the performance of our system via CMC curve analysis. More specifically, each sequence in the probe is tested against the training set and the ranked list of Re-ID is obtained via signature matching. (The rank is computed by person i.e. best of the three sequences.) The process is repeated for all probe sequences. Then the average over all probe sequences Re-ID is computed and represented as CMC result.

5.3.1 Experiment 1: View-point dependent Re-ID

In this experiment, we test Re-ID in individual directions. This is done to verify the performance of the proposed method along specific directions. Or in other words, we test how well the system can act when both the probe and gallery contain the features extracted in a particular direction. We carry out a leave-one-out evaluation strategy, in which any of the gait sequences will be selected as a probe and tested against the remaining 59 sequences. This is then repeated 60 times, with each of the gait sequence used exactly once as the test data, and the average Re-ID result is computed.

We exploit both the anthropometric and gait features. Regarding the anthropometric features, we select seven body measurements: *height, arm length, upper torso, lower torso, upper-lower ratio, chest and hip* (see Table 5.1 for the list of features). An example for the estimation of 'height' feature is shown in Fig. 5.6, by calculating the mean information within a gait period.

First, we analysed the Re-ID ability of our framework exploiting individual features. An example of CMC curve produced from each anthropometric features in frontal view is shown in Fig. 5.7(a). Among them, the most informative features are the height and arm length information with Rank-1 CMC accuracy of 65.9% and 48.7% respectively.

Similarly, we also analysed the impact of other individual gait features separately. Please refer to Fig. 5.7(b). It includes various body angles, distances and evolution of certain joints, along the time. The mean and variance information are extracted to generate the feature vector. We noticed that, all of those gait features are less informative and distinguishable in comparison with the anthropometric features. Refering to Fig. 5.7(b), the important gait features are the elbow distance and hand distance achieving Rank-1 CMC rates 51.67% and 30%, respectively whereas the least informative features were speed and stride length which achieved 5.12% and 2.5% accuracy respectively.

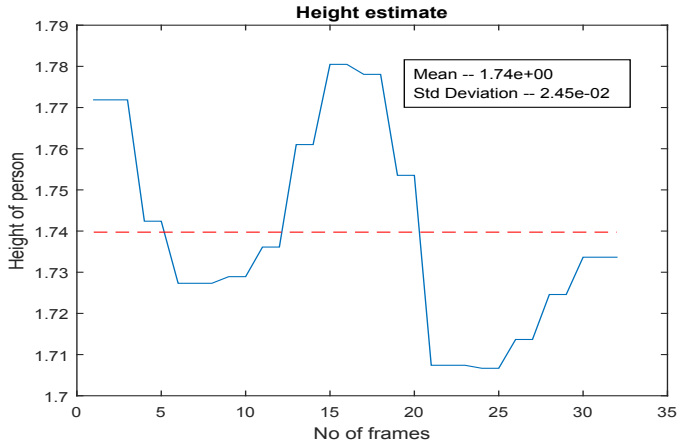


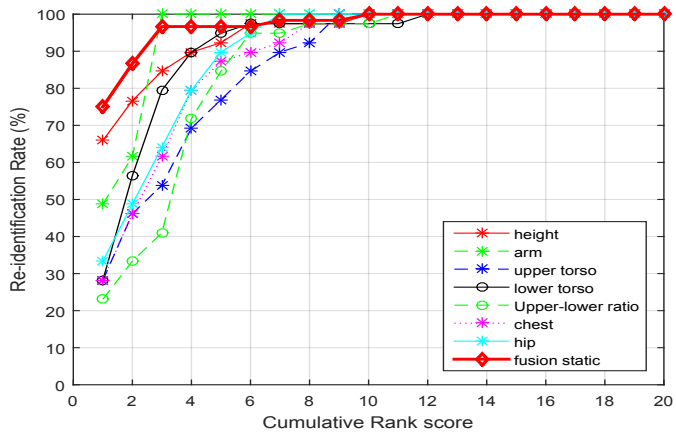
Figure 5.6: Height estimation from the sequence of frames within a gait cycle.

Next, we conducted fusion of the multiple features aka multi-modal fusion. Initially, various anthropometric features were fused together which resulted in the bold red CMC curve in Fig. 5.7(a), which achieved 75% Re-ID rate at Rank-1. Similarly, the fusion of gait features were also conducted. The result is shown with the bold blue CMC curve in Fig. 5.7 (b), which achieved 61.67% Rank-1 Re-ID rate. We could observe that, fusion of body related measurements produced higher Re-ID performance in comparison with the fusion of the gait features. It was quite noteworthy that even by combining 35 gait features, it couldn't achieve similar Re-ID accuracy as obtained by the anthropometric fusion by seven features. This gives the intuition that in frontal view, anthropometrics features are more significant than the gait features in discriminating the population.

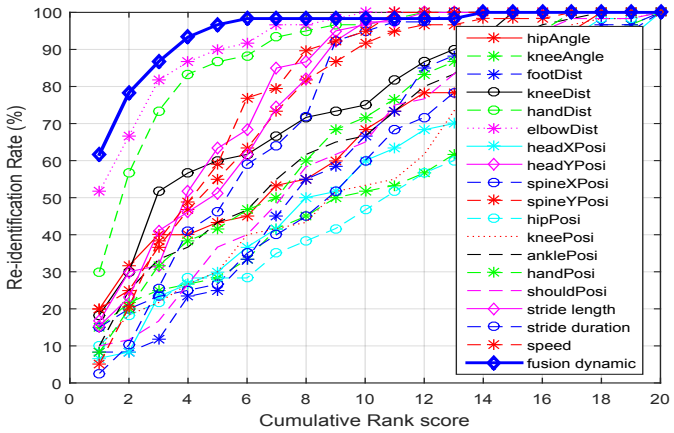
After conducting the fusion among the anthropometric features and gait features separately, we further conducted the multimodal fusion of all the biometric features (i.e. both anthropometric and gait features), altogether. The results obtained in these multi-modal fusion technique in frontal sequence is presented together in Fig. 5.8(a). Red and blue curves denote anthropometric fusion (75% Rank-1 score) and gait fusion (61.67% Rank-1 score) result respectively. The combined anthropometric+ gait fusion result is represented via green curve with a Rank-1 Re-ID accuracy of 91.67%. We could observe that the naïve integration could improve the overall performance while fusing both anthropometrics and gait features together.

Similar experiments are also conducted in the other three views as well, i.e. left diagonal, right diagonal and lateral. We show the fusion results of all the three experiments in Fig. 5.8(b), (c) and (d) with an overall Rank-1 scores of 71.67%, 63.33% and 70%, respectively. In all these scenarios also, we could observe that the anthropometric features outperform the gait features. Also, while fusing both the anthropometric and gait features together, the overall performance improved.

A similar human classification strategy based on gait features has been reported in [Gianaria *et al.* , 2014], by employing 20 people. In contrast to our methodology, they have conducted the experiments only in a single view (i.e. frontal) as well as an exhaustive se-



(a)

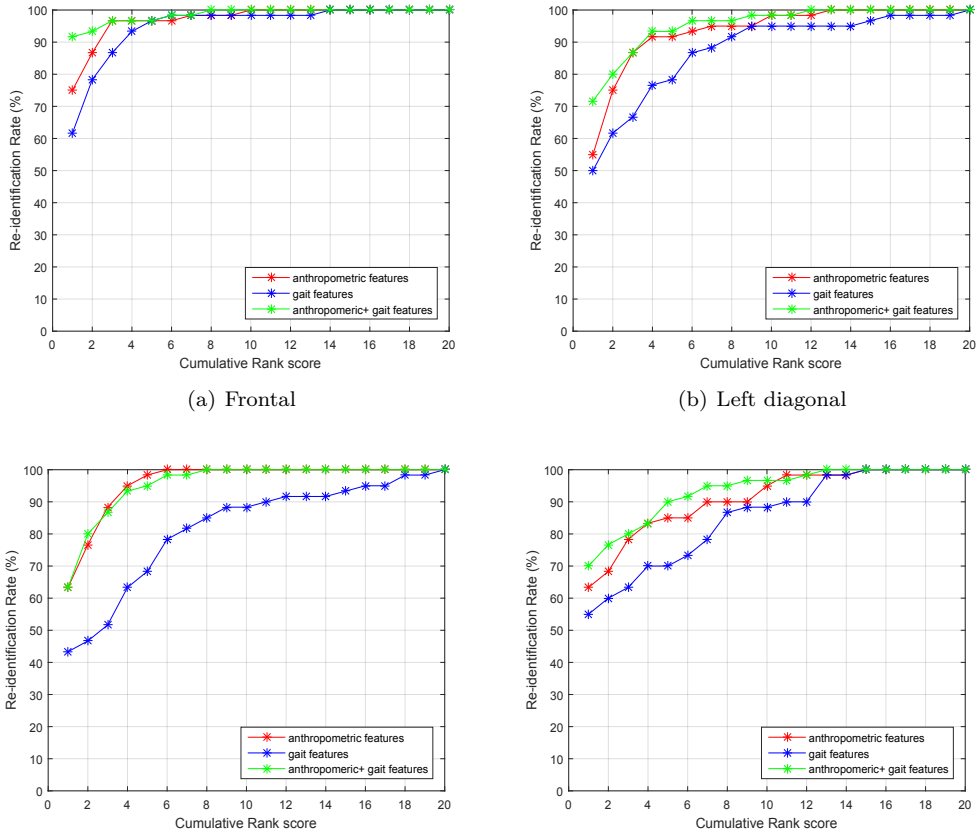


(b)

Figure 5.7: Individual feature performance towards Re-ID: (a) Static anthropometric features and scalar gait features (stride length, stride time and speed). The bold red curve with diamond markers corresponds to the fusion CMC result obtained by exploiting all the anthropometric features. (b) Dynamic gait features. The result by fusing all the gait features is shown in bold blue curve with diamond markers.

lection of the set of different features along with a SVM classification scheme. However, our experiments were explicitly made in different views, and via naive score-level fusion of multi-modalities. Hence, an approximate comparative analysis is made at this point, particularly Fig. 5.8(a) referring to the frontal Re-ID experiment. The highest classification accuracy observed in their case is 96.25% (19.25 times the chance level[7]) under fine tuned parameter set (elbow distance, knee distance, mean of head, mean of knee). Nevertheless, our direct approach of naive fusion also could achieve quite similar result 91.67% (18.34 times the chance level) without the exhaustive feature search or the fine tuning of the parameter set.

[7]Chance level is Re-ID of 1 subject out of 20 subjects, i.e. 0.05.



(a) Frontal (b) Left diagonal

(c) Right diagonal (d) Lateral

Figure 5.8: Multimodal fusion of anthropometric features, gait features (using mean-variance) and the fusion of both, in various directions. (a) Frontal (b) Left diagonal (c) Right diagonal (d) Lateral.

5.3.2 Experiment 2: View-point independent Re-ID

In Section 5.3.1, we have conducted experiments along various view angles at ∼0°, ∼30°, ∼60° and ∼90°, separately. Albeit we could analyse the impact of various features in each of these directions, we did not so far experiment how feasible and robust is our system in order to perform in view-point invariant scenario i.e. irrespective of any particular direction. Hence, we conduct a thorough analysis of various view-point independent Re-ID schemes i.e. pseudo view-point invariant, quasi view-point invariant and full view-point invariant.

Pseudo view-point invariant Re-ID experiment:

In pseudo view-point invariant case, we consider that the gallery contains samples from multiple views. And, the probe will be a new sample taken from any of these views. This kind of set up requires either a large number of cameras with different camera views (in the case of normal surveillance case), or the person's different views acquired in the enrollment phase (authentication phase). The nomenclature 'pseudo' is attributed to the fact that the probe view is already encountered among the gallery views and hence its a pseudo view-point invariant

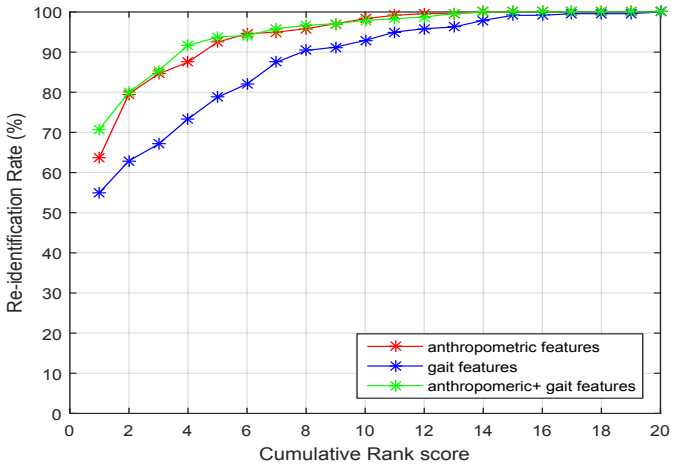


Figure 5.9: Pseudo view-point invariant Re-ID results using anthropometrics+ gait.

Re-ID.

Since we have used 20 people's gait in four different directions, each with three sequences, altogether we have 240 gait sequences. We conduct a leave-one-out evaluation strategy, in which any of these sequences will be selected as a probe and tested against the remaining 239 sequences in different views. Altogether 240 runs were conducted and the averaged result was computed. The achieved performance of the system is depicted in Fig. 5.9.

We could observe that, the fusion of anthropometric features achieved 63.75% (red curve in Fig. 5.9) and the fusion of gait features achieved 55% with (blue curve in Fig. 5.9) respectively. While combining both of them, we could obtain improvements in their performance i.e. ∼71% Rank-1 Re-ID rate. This is a promising result highlighting the performance and robustness of our system towards handling various direction of gait, which is a big challenge in the Re-ID task. Our intuition is that the increased number of samples per person (12 sequences) compared to a single direction (three sequences) could enhance the Re-ID rate.

Quasi view-point invariant Re-ID experiment:

Here, in the quasi-view-point invariant scenario, the gallery contains multiview samples of the subjects. However, the probe sample is taken from a new view angle which has not been introduced in the training phase. This is a realistic scenario, where a new camera view is encountered in which the person has to be re-identified, provided that many other training samples in different views are available in the gallery. This is a more challenging case than the pseudo view-point invariant case, since the probe direction is encountered in the system for the first time.

In order to test this case, we keep all the samples in a particular direction in the test set, whereas all the other three directions are made available in the training phase. In particular, we have 180 gait sequences of 20 people corresponding to three directions being kept in the training set. The 60 gait sequences from the fourth walking direction (which was not introduced in the training phase) are used for testing. Hence, 60 runs per view are carried out and the

average result is estimated. We conduct the experiment for all the frontal, left diagonal, right diagonal and lateral views as the test direction.

The Re-ID rates at Rank-1, Rank-5 and Rank-10 are presented in Table 5.3. It is observed that the highest Rank-1 CMC rate for the anthropometric fusion is reported in the frontal view case (41.33%) and the counterpart for the gait fusion was reported in lateral view (31.67%). Coherent results were also observed in the fusion of anthropometric+ gait case as well, where frontal samples got re-identified with the highest recognition rate (65%) followed by lateral samples (41.67%) among all the directions, in the Rank-1 scenario. With Rank-5 and Rank-10 rates in CMC curves, the Re-ID accuracy improved drastically >73.33% and >90% respectively, in all the directions. Once again the highest Re-ID rates were reported in frontal case (Rank 5- 86.67% and Rank 10- 98.33%). This means that, given other multiple views in the gallery set, frontal view probes are the best in re-identifying people.

Table 5.3: Chart showing the Re-ID accuracy rates for Experiment 4.2.2. The accuracy rates shown in each cell represents Rank-1, Rank-5 and Rank-10 CMC rates respectively. The highest Re-ID rate observed is highlighted in bold letters.

Probe direction	**Anthropometric based Re-ID**	**Gait based Re-ID**	**Anthropometric + gait based Re-ID**
Frontal	**41.33%** **90.00%** **98.33%**	26.67% 68.33% 96.67%	**65.00%** **86.67%** **98.33%**
Left Diagonal	33.33% 73.33% 91.67%	21.67% 53.33% 88.33%	28.33% 73.33% 90.00%
Right Diagonal	28.33% 80.00% 93.33%	10.00% 56.67% 90.00%	31.67% 83.33% 93.33%
Lateral	40.00% 68.33% 93.33%	**31.67%** **70.00%** **81.67%**	41.67% 75.00% 96.67%

Full view-point invariant Re-ID experiment:

Full view-point invariance is the case which has only one walking direction in the gallery and any new arbitrary walking direction for the probe. In terms of creating a training set, this is the easiest way because it requires only one camera and one view of the person to create a gallery. At the same time, it is the most challenging scenario in terms of Re-ID, since it requires to get recordings from merely one view and able to Re-ID in any other arbitrary view.

We conducted 12 various combinations of probe-gallery set based on the walking direction, in order to guarantee a truly view-point invariant Re-ID. The experiments and the results achieved are reported in Table. 5.4. In each of the test case (e.g. frontal), we keep any of the other three view-point data sequences as the gallery (e.g. left diagonal or right diagonal or lateral). And the same procedure is repeated for all the four directions. In all of these experiments, each of the probe and gallery contains 60 gait sequences from 20 people. Per each combination, 60 runs were carried out and the average Re-ID resut is estimated. In the tabular results (see Table. 5.4), we report only the overall anthropometric+ gait multimodal fusion results at various ranks (Rank-1, 5 and 10) of CMC curves. It is observed that the highest Re-ID rates (48.33%) are achieved when frontal sequences are kept in the gallery. With the diagonal samples the second best Re-ID results are achieved (∼35%).

Despite most works assume that kinect data is pose invariant, this is not really the case as demonstrated in all the experiments of our work. Re-ID rates are always better in the frontal view that in the other, due to the quality of the data acquired. We show that with an adequate use of pre-processing and soft biometrics we can achieve some level of view-point invariance, but still not perfect.

Table 5.4: Chart showing the Re-ID accuracy rates for Experiment 4.2.3. The accuracy rates shown in each cell represents Rank-1, Rank-5 and Rank-10 CMC rates respectively. The highest Re-ID rate observed is highlighted in bold letters.

		PROBE			
		Frontal	Left Diagonal	Right Diagonal	Lateral
GALLERY	Frontal	-	26.67%	**48.33%**	**48.33%**
		-	78.33%	88.33%	73.33%
		-	91.67%	93.33%	93.33%
	Left Diagonal	33.33%	-	30.00%	35.00%
		75.00%	-	70.00%	78.33%
		90.00%	-	85.00%	96.67%
	RightDiagonal	35.00%	25.00%	-	18.33%
		85.00%	68.33%	-	58.33%
		95.00%	83.33%	-	85.00%
	Lateral	18.33%	28.33%	15.00%	-
		78.33%	78.33%	68.33%	-
		90.00%	93.33%	86.67%	-

5.4 Summary

A view-point invariant Re-ID system exploiting the skeleton information provided by the kinect sensor has been proposed. We have used both the static and dynamic features related to the human posture and walking, in order to extract features to classify the people in the population. Extensive study on the impact of various features both individually and jointly, as well as various view angles have been conducted. We have acquired the kinect data in-house from 20 people walking in four different directions, and analysed our proposed methodology.

We could observe that the static anthropometric features are more informative than gait features, when employed individually. However, while fusing many static anthropometric features and dynamic gait features, we noticed that the overall recognition accuracy increases in both cases. Also, by combining the whole set of static and dynamic features, the final overall Re-ID rate improved further. In addition to evaluations in individual directions, we also conducted view-point invariant Re-ID experiments in realistic conditions where people walk in different directions. Three cases studies were conducted in this regard viz. *pseudo, quasi* and *full* view-point invariant. It is found that our system is quite robust and promising with a Rank-1 Re-ID rate of ∼92% in view-point dependent scenarios and ∼71%, ∼65% and ∼48% in pseudo, quasi and full view-point independent scenarios, respectively. Since the direct comparison with other works are not possible due to the novelty of the approach, we carry out comparative analysis against the most similar view-point dependent approach [Gianaria *et al.* , 2014] in the front view, and very similar Re-ID results (19 times and 18 times the chance level, respectively) were reported.

In the future, we envisage to extrapolate this study by collecting more data in more random directions of walk. Also, in terms of the feature fusion, we would like to employ context based

fusion or feature selection strategies (eg: quasi-exhaustive learning strategy [Barbosa *et al.* , 2012b], correlation-based feature subset selection [Andersson & Araujo, 2015]), in order to fine tune the selection of most informative features and thus improve the Re-ID accuracy.

Chapter 6

Context-Aware Person Re-ID

Content is King, But Context is God.

— Gary Vaynerchuk

6.1 Introduction

Soft-biometric enabled feature extraction depend strongly on the view-point. For instance, a person with a short stride gait is better perceived from a lateral view, whereas a person with a large chest is more distinct from a frontal view. Thus we associate context to the viewing direction of walking people in a surveillance scenario and choose the best features for each case.

In this chapter, we discuss the application of soft-biometrics (anthropometrics and gait) in long term Person re-identification (Re-ID) using cameras in arbitrary view-points. We study the influence of the view-point in Re-ID performance and propose a methodology to exploit the *'view-point context'* to improve the overall performance. The best features for each context are selected for training context specific classifiers. Then, during run time, a context-specific fusion method provides the person Re-ID score. Based on these concepts, we present a novel 'Context-aware ensemble fusion Re-ID framework' based on soft-biometric features, for long term person re-identification (Re-ID) in wild surveillance scenarios.

Some works have employed Kinect based person Re-ID approaches leveraging soft-biometric cues [Barbosa *et al.* , 2012b; Gianaria *et al.* , 2014; Andersson & Araujo, 2015]. Nevertheless, they employed view-point dependent methods *i.e.*, data was collected and algorithms were tested with a single walking direction with respect to the camera, which does not represent a 'in the Wild' scenario where people walk in various directions. On the contrary, in this work, we collect people walking freely in an indoor office like scenario. Depending upon the strategic points inside a building (entry/exit points, and coffee machine/printer locations etc.), it was observed that the probability of people walking indoor could be explicitly represented in various directional view-points, which we term as '*Contexts*', rather than random walking paths. In addition to that, the potential features extracted by the sensor also have indicated clear distinction, according to different contexts. Based on these postulates, we redefine the classical Re-ID strategy by means of a novel 'context-aware person re-identification method', where we explicitly evaluate a context-specific feature matching criteria in Re-ID.

In this regard, the major contributions of the paper are as follows:

- We propose to consider view-points as 'contexts' for re-identification (Re-ID) systems based on the fact that the features that best correlate with peoples' identity depend strongly on the view-points.

- Model each context with a specific set of features selected with Sequential Forward Selection (SFS) algorithm, to adaptively select the potentially relevant features in each context and thus to maximize Re-ID score in each context.

- Proposal of a '*Context-aware ensemble fusion framework*', wherein individual classifiers are trained specific to each context, and the Re-ID performance is analysed via our proposed '*Context-specific score level fusion*' strategy.

- Proposal of a new dataset with 20 people walking in 5 different directions acquired from Kinect v.2, suitable for pose-invariant Re-ID.

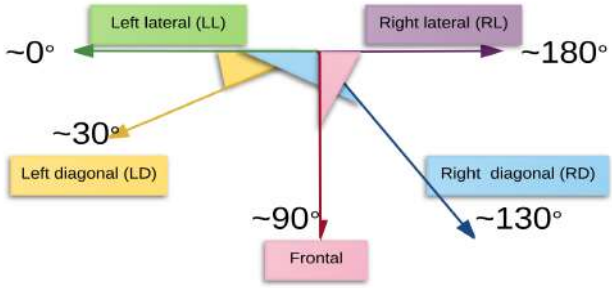
This chapter is organized as follows. The proposed methodology is explained in Section 6.2, *i.e.*, the dataset used, feature extraction method, and Context-aware ensemble fusion framework. In Section 6.3, the experiments conducted and the results obtained are discussed in detail. Finally, the summary of the paper and some future plans are enumerated in Section 6.4.

6.2 Methodology

6.2.1 Database

In order to employ Re-ID in a realistic 'in-the-wild' scenario, it is quite essential to have a challenging unconstrained dataset, comprised of sequences of people walking in different directions. Since such a Kinect™ based dataset (with different viewangles) towards gait based Re-ID was unavailable, we acquired our own dataset using a mobile platform, in the host laboratory. The Kinect™ device is able to track movements from users by using a skeleton mapping algorithm, and is able to provide the 3D information related to the movements of body joints. The position of camera as well as the walking directions of subjects were deliberately altered in order to ensure a typical surveillance scenario. Multiple walking sequences of 20 subjects in five different directions *i.e.*, Left Lateral (LL) (at ∼0°), Left Diagonal (LD) (at ∼30°), Frontal (F) (at ∼90°), Right Diagonal (RD) (at ∼130°) and Right Lateral (RL) (at ∼180°) were collected. Altogether we have 300 video sequences comprising 20 subjects (3 video sequences per person in a particular context) in the aforementioned directions. Different walking directions and sample video frames extracted from our dataset, are shown in Fig. 6.1. Later on, we released this dataset to the community for research purpose under the name **'KS20 VisLab Multi-View Kinect skeleton dataset'**[1].

[1]KS20 VisLab Multi-View Kinect skeleton dataset is made publicly available from May 2017 onwards. The link of the website is http://vislab.isr.ist.utl.pt/vislab_multiview_ks20/. Access to the Vislab Multiview KS20 dataset is available upon request.



(a)

(b) (c) (d) (e) (f)

Figure 6.1: Data acquisition: (a) Subject walking directions in front of the camera system (Direction angles are defined with respect to the image plane.) (b-e) Sample frames from our data acquisition, in five different directions- left lateral (∼0°), left diagonal (∼30°), frontal (∼90°), right diagonal (∼130°) and right lateral (∼180°) respectively.

6.2.2 Feature extraction

The real-time skeleton models tracked via Kinect^{TM} are composed of 25 body joints. The foremost step in feature analysis was preprocessing, to remove the noise contents in the data. By empirically analysing the evolution of lower body angles over time, we cleared the unwanted jerks in the signals especially, at the boundaries of the Kinect range. The detailed explanation of pre-processing and feature extraction phases were reported in the prior work by the authors in **Chapter5** and in [Nambiar *et al.* , 2017b] [2]. Then, based on those cleaned signals, the functional units of gait *viz.*, gait cycles, were estimated. A gait cycle comprises of sequence of events/movements during locomotion since one foot contacts the ground until the same foot again contacts the ground. Hence, based on the cleaned data, the periodicity of the feet movement is estimated to define gait cycle and various features were extracted within this gait period.

Two kinds of features were extracted: *(i)* Anthropometric features *i.e.*, the static physical features defining the body measurements and *(ii)* Gait features *i.e.*, dynamic features defining the kinematics in walking. See Table 6.1 for the list of features we used. Under the anthropometric feature set, body measurements defining the holistic body proportions of the subject such as height, arm length, upper torso length, lower torso length, upper to lower ratio, chest size, hip size were collected. Similarly, under the gait features, the behavioural features deriving from the continuous monitoring of joints during the gait were collected. In particular, mean and standard deviation of the various measurements during a gait cycle were collected

[2]The key idea of the prior work was to analyse the influence of various anthropometric and gait features either individually or jointly (via fusion), and to demonstrate the real impact of view-point on the Re-ID paradigm. The results highlighted (i) the significance of multi-modal fusion strategy in overall Re-ID results, and (ii) not all the features are equally contributing towards various view-points, *i.e.,* the Re-ID result varied in different view-points (although by using the same features). In the current paper we extend that prior work by adding the *i.e.*, Feature selection strategy and Context-aware fusion framework.

Table 6.1: List of anthropometric and gait features used in our experiments. L& R correspond to 'left and right' and x& y correspond to 'along x and y axes'. The numbers of features derived are shown within parenthesis.

Anthropometric features	**Gait features**	
Height-(*1*)	Hip angle(L&R)-(*4*)	Hip position(L&R)(x& y)-(*8*)
Arm length-(*1*)	Knee angle(L& R)-(*4*)	Knee position(L&R)(x& y)-(*8*)
Upper torso-(*1*)	Foot distance-(*2*)	Ankle position(L&R)(x& y)-(*8*)
Lower torso-(*1*)	Knee distance-(*2*)	Hand position(L&R)(x& y)-(*8*)
Upper-lower ratio-(*1*)	Hand distance-(*2*)	Shoulder position(L&R)(x& y)-(*8*)
Chestsize-(*1*)	Elbow distance-(*2*)	Stride-(*1*)
Hipsize-(*1*)	Head position(x& y)-(*4*)	Stride length-(*1*)
	Spine position(x& y)-(*4*)	Speed-(*1*)

i.e., (i) the angles at various body joints; (ii) the distance between various right-left limbs and; (iii) the position of body joints. Also three scalar features related to walking, *viz.*, stride length, stride time and the speed of walking, are computed within the gait features. Hence, the feature set contains a total of 7 anthropometric features and 67 gait features. In Table 6.1, the numbers of features derived are shown in parenthesis.

6.2.3 Context-aware ensemble fusion

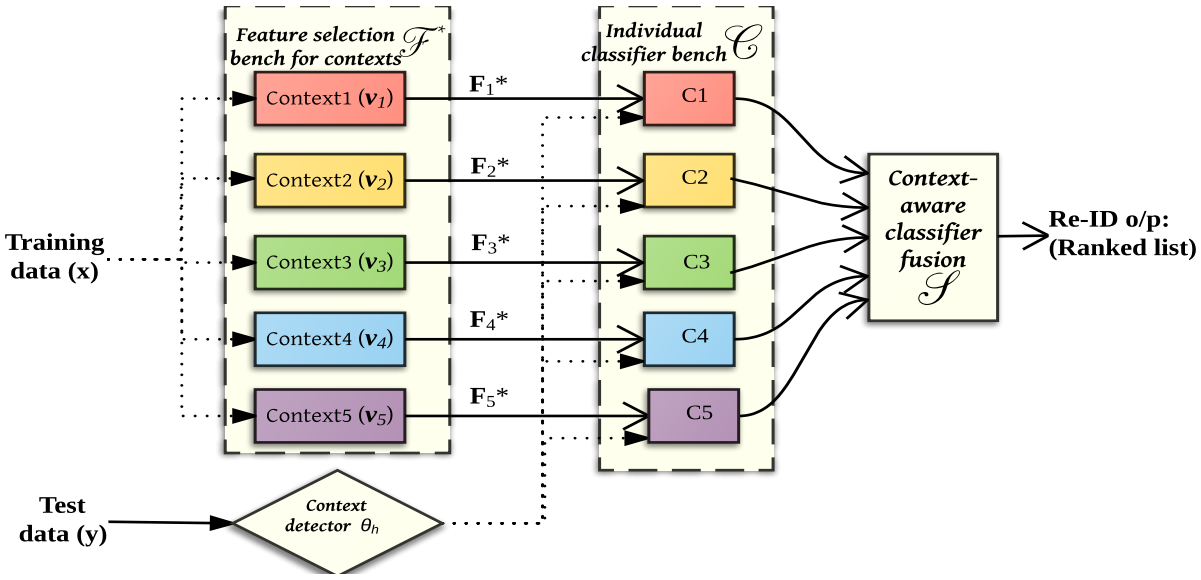


Figure 6.2: Context-aware ensemble fusion system: It internally consists of a feature selection context bench, an individual classifier bench, a classifier fusion module and a context detector module. The individual classifiers for each context are trained using individual feature subspace ensembles $\mathbf{F}_\mathbf{j}^*$, obtained for each context. When the test data enters, context detector identifies the context and activates the corresponding ensemble classifiers. Then, the context-aware classifier fusion strategy finally combines the results of those ensemble classifiers to produce the global result.

One of the most significant contributions of this work is a novel context-aware ensemble fusion strategy. First, we present an evaluation of the impact of the various data features in various contexts *i.e.*, view-points, and then employ a context-based fusion method to obtain the final Re-ID result. We accredit the work on Feature subspace ensembles [Silva & Fred,

2007] which acted as a motivation to the authors to come up with an analogous ensemble fusion strategy. That work presented an approach to run multiple parallel Feature selection stages with different training conditions, in order to obtain the best features, by using majority voting of the feature ensembles.

Our proposed framework is shown in Fig. 6.2. It is composed of four modules: (i) *Feature selection Context bench* (ii) *Individual classifier bench*, (iii) *Context detector module* and (iv) *Context-aware classifier fusion module.*

Feature selection Context bench

Our data for evaluation consists of the feature vectors extracted at various view-points, as mentioned earlier. We denote those five context view-points as $\mathbf{v_1}, ..., \mathbf{v_N}$, with $N = 5$, corresponding to LL, LD, F, RD and RL directions. We analyse the data in each context individually by leveraging a Feature Selection (FS) scheme in order to retain only the most discriminative and relevant features.

In particular, we employed Sequential Forward Selection (SFS) algorithm [Whitney, 1971] as an instance of FS, as it is well known and widely used in practice. It works iteratively by adding features to an initial subset, seeking to improve a given measure, by selecting more features at each iteration. Suppose, $\mathbf{x} = \{x_1, \cdots, x_n\}$ denotes a set of n samples represented in a d-dimensional space, each with a d-dimensional feature set $\mathbf{F} = [f_1, \cdots, f_d] \in \mathbb{R}^{1\times d}$. FS analyses this d-dimensional space in order to identify which features $f_i \subset \mathbf{F}$ are potentially relevant, and which can be discarded according to some feature subspace evaluation criteria $\boldsymbol{J}$ and ultimately derive $\mathbf{F_j^*}$, containing the most relevant features.

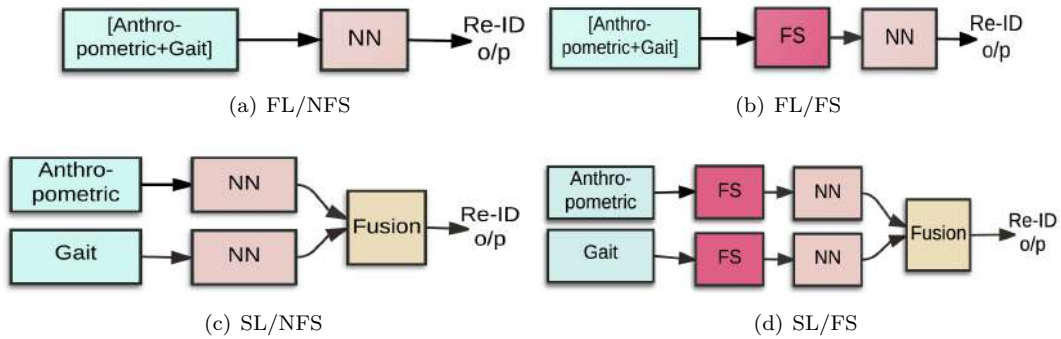
Specifically, the SFS algorithm works as follows: It starts from an empty feature set $\mathbf{F^*_{t=0}}$. At each step $\mathbf{F^*_{t+1}}$ all possible super-spaces containing the most relevant feature subspace in the previous step, $\mathbf{F_t^*}$, and one from the remaining features $f_i \in \mathbf{F}\backslash\mathbf{F_t^*}$ are formed and evaluated by $\boldsymbol{J}$. This iterative search will proceed until a stopping criteria is met, for which we considered the degradation of $\boldsymbol{J}$ *i.e.*, if none of the super-spaces formed at a given step $\mathbf{F^*_{t+1}}$ improves $\boldsymbol{J}$, the search stops and the subspace $\mathbf{F_t^*}$ is considered as the best feature subset. Finally, the outputs of the Feature selection context bench consists of an ensemble of feature subspace *i.e.*, the features selected for each particular context $\mathcal{F}^* = [\mathbf{F_1^*}, \cdots, \mathbf{F_5^*}]$. For the implementation of the algorithm, the authors used SFS package[3] Pohjalainen *et al.* [2015]. We used 1-NN classifier with an Euclidean neighborhood metric in the SFS scheme.

Individual classifier bench

Since our training data consists of both anthropometric and gait features, we need to exploit both of them in training our each individual classifier. In this regard, we exploit various fusion techniques in order to combine anthropometric and gait features. Traditionally, there are many fusion strategies at various levels *viz.*, feature level fusion, score level fusion, rank level fusion or decision level fusion [Ross *et al.* , 2006], of which we select both feature level fusion and score level fusion strategies in our work. In order to see the impact of various fusion strategies, we conduct two baseline fusion schemes without Feature selection: (i) Feature-level fusion without FS, represented as Feature Level fusion with No Feature Selection (FL/NFS) and (ii) Score-level

[3] http://users.spa.aalto.fi/jpohjala/featureselection/

fusion without FS, represented as Score Level fusion with No Feature Selection (SL/NFS). The schematic representations of the aforementioned are shown in Fig. 6.3 (a) FL/NFS and (c) SL/NFS respectively.



(a) FL/NFS (b) FL/FS

(c) SL/NFS (d) SL/FS

Figure 6.3: Various Fusion-Feature selection schemes employed in this work. Top and bottom rows represents feature-level and score-level fusion strategies respectively. Feature Selection (FS) is not used in case studies (a) FL/NFS and (c) SL/NFS, whereas (b) FL/FS and (d) SL/FS shows the inclusion of FS module.

In Feature level fusion (see Fig. 6.3 (a)), the biometric sets of the same individual are concatenated after an initial normalization (Min-max) scheme. This way, we concatenate our 7D anthropometric features and 67D gait features in order to make a 74D feature vector. Then, the concatenated feature vector is used in the classifier in order to represent the identify of an individual. Instead, in score level fusion (see Fig. 6.3 (c)), the fusion is carried out at the score level. The matching scores of each biometric sets are determined independently using two different classifiers and the matching scores at their outputs are fused in order to provide an aggregate score result. As explained in [Ross *et al.* , 2006], normalized distance scores obtained at each individual classifiers can be fused using some combination rule such as sum, product, min, max or median. In our approach, we adopted sum rule as the classifier combination rule.

After the baseline cases, we further conduct our proposed FS-enabled fusion strategies as well. Here, the biometric sets are fed into a FS module prior to the classification stage so that, only the selective feature subspace $\mathbf{F}_\mathbf{j}^*$ (as explained in Section 6.2.3) will be used as the individual feature vector. In this regard, two more fusion schemes with Feature selection are carried out: (i) Feature-level fusion with FS, represented as FL/FS and (ii) Score-level fusion with FS, represented as SL/FS. The schematic representations of the aforementioned are shown in Fig. 6.3 (b) FL/FS and (d) SL/FS respectively.

Thus, as explained above, four different Fusion-FS schemes are conducted in order to assess the performance of each individual context classifiers within the classifier bench. In all of those case studies, a leave one out evaluation strategy is performed within each context, with a classifier specification of Nearest neighbour (NN) using euclidean distance metric. The experimental results obtained are explained in Section 6.3.1, and the best among all those fusion-FS scheme is further used as the *de facto* standard scheme in our framework. Based on this standard scheme, five different classifiers are trained corresponding to each context, which will form the Individual Classifier bench $\mathcal{C} = [\mathbf{C_1}, \cdots, \mathbf{C_5}]$.

Context detector

Context detector is the module where the context (view-point) of the test sample is estimated. The design of the context detector module was carried out by analysing the evolution of any static joint along the sequences over a gait cycle. We used 'SpineShoulder' *i.e.*, the base of the neck refering to joint number 20 of Kinect^{TM} v.2[4], since it remains more or less stable while walking. Then, the direction of walking was estimated by analysing the direction of the joint vector. Suppose $h_{\vec{begin}}$ and $h_{\vec{end}}$ denotes the position of the joint in the first frame and last frame respectively. Then the directional vector among these frames $\vec{h} =< h_x, h_y, h_z >$ can be obtained as follows:

$$\vec{h} = h_{\vec{end}} - h_{\vec{begin}}, \tag{6.1}$$

The y component h_y is only related to the vertical direction and hence is ignored. Then, the angular direction $\theta_{\vec{h}}$ made by $\vec{h}$ can be determined by measuring the inverse tangent of h_z/h_x.

$$\theta_{\vec{h}}(degrees) = \tan^{-1}(h_z/h_x) * 180/\pi \tag{6.2}$$

Whenever a test data $\mathbf{y} \in \mathbb{R}^{1\times d}$ enters into the system, its context is estimated using (6.1) and (6.2), and the corresponding ensemble classifiers are activated in order to proceed with context-aware classifier fusion.

Context-aware Classifier fusion

Based on the results from context detector module, this classifier fusion module performs a context-specific adaptive fusion of the results obtained at the outputs of individual classifiers $\mathcal{C} = [\mathbf{C_1}, \cdots, \mathbf{C_5}]$. In order to facilitate this, an extended version of score-level fusion based on context is proposed in this work, which we term as '***Context-specific score level fusion***'. This could be analysed homologous to the concept of user-specific score-level fusion in multibiometric systems, where user-specific weights were assigned to indicate importance of individual biometric matchers [Ross *et al.* , 2006]. In a similar way, in our proposal, we endorse adaptive weights to scores from different classifiers according to its context, in order to increase the influence of more reliable context. In order to facilitate this adaptive weighting scheme, we employ linear interpolation technique.

Consider a test sample $\mathbf{y}$, at an arbitrary view-point context $\mathbf{v_{test}}$, is entering into the system. The context is detected using the context-detector module. Suppose the context lies in between our pre-defined context views say, $\mathbf{v_i}$ and $\mathbf{v_j}$. The individual classifiers for both aforementioned contexts $\mathbf{C_i}$ and $\mathbf{C_j}$ are selected alongwith their matching scores $\mathbf{s_i}$ and $\mathbf{s_j}$ respectively. The context-specific score level fusion $\mathcal{S}$ is computed as weighted sum of those scores as follows:

$$\mathcal{S} = \eta * \mathbf{s_i} + (1 - \eta) * \mathbf{s_j}, \tag{6.3}$$

where $\eta \in [0, 1]$. The weight η is computed via linear interpolation of the two contexts *i.e.*, $\eta = |\mathbf{v_j} - \mathbf{v_{test}}|/|\mathbf{v_j} - \mathbf{v_i}|$. The special case where only single context is activated, η of the

[4] https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx

nearest context turns to be 1, and all the others will be 0. Regarding these concepts, we analyse different case studies in detail, in the experimental section Section 6.3.2.

6.3 Experimental results

The performance of the context-aware ensemble fusion strategy was evaluated on our own database collected from 20 people, mentioned in previous section (see Section 6.2.1). Two major experiments were carried out: *(A) Training the individual context-specific classifier*, in which each individual classifier was learned specific to its context. Intermediate experiments leading to this standard scheme (such as various fusion-FS schemes for performance assessment, context-specific feature ensembles selected) are also detailed in this section. The second experiment is *(B) Context-Specific Score Level Fusion*, wherein the final Re-ID result was achieved via adaptive fusion of the ensemble classifiers. Under this part, the experiments on context detection and context-specific fusion strategy are detailed. In order to evaluate the performance of our Re-ID algorithms, we use the popular method of choice, cumulative matching characteristic (CMC) curve. As per [Nambiar *et al.* , 2014b], "CMC shows how often, on average, the correct person ID is included in the best K matches against the training set for each test image".

6.3.1 Training the individual context-specific classifiers

In this step, we assessed the individual context classifier performance leveraging the 7D anthropometric features and 67D gait features. The impact of various fusion and FS schemes were analysed in this stage, via the four extensive case studies explained in Section 6.2.3.

Why Feature selection is required?

Prior to the selection of feature subspace ensembles, initially we tried to analyse the Re-ID performance of each gait as well as anthropometric feature individually. This is performed with the NN classifier explained before but using a single feature *i.e.*, no fusion. Fig. 6.4 shows the individual performances of some of the best features[5] in person Re-ID in the frontal context. We can observe that certain features are quite relevant and discriminative (*e.g.*, height 55.89%, arm length 41.67%, elbow distance 50.00% and chest size 35.00%) compared to the others in re-identifying people. Another interesting result was that the feature level fusion of all features among a biometric set (gait or anthropometric) resulted in better performance. Or in other words, multi-modal fusion outperformed the individual Re-ID results. Referring to the CMC curves in Fig. 6.4, we can observe that fusion of all anthropometric features resulted in 85.00% Re-ID rate at Rank-1 (bold green curve) and fusion of all gait features resulted in 60.00% Re-ID rate at Rank-1 (bold blue curve).

After the feature level fusion of anthropometric features and gait features separately, we further conducted multimodal fusion of the biometric features altogether *i.e.*, both anthropometric and gait features. Within this scheme, we first utilized feature-level fusion strategy *i.e.* as per Fig. 6.3 (a) FL/NFS. However, we could observe that the multimodal fusion at feature

[5] Among 74 features, only those features with individual Re-ID performance $\geq$ 20% are illustrated here.

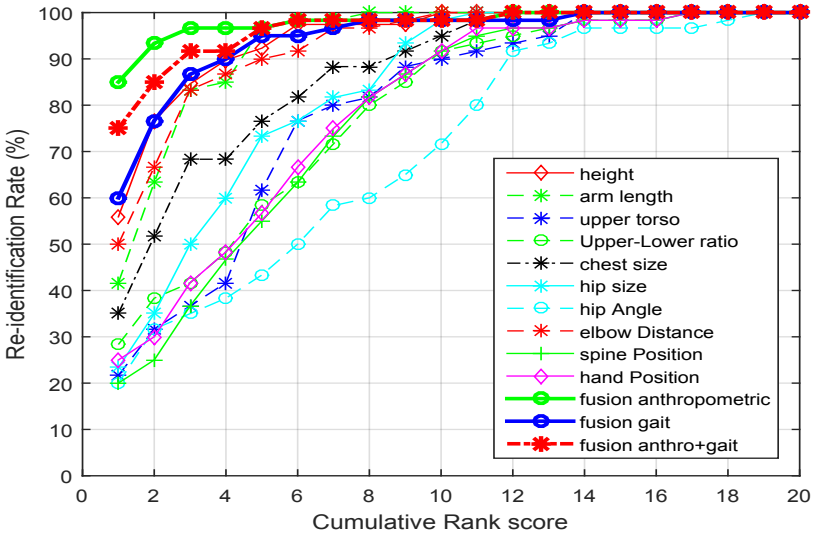


Figure 6.4: Re-ID performances of individual as well as fused features in frontal context. Only a subset of individual features with classification rate $\geq 20\%$ at Rank-1, are shown. The fusion results of anthropometric features (green with circle markers), gait features (blue with circle markers) and both anthropomeric and gait features (red with star markers) are shown via bold curves. For fusion of features, feature-level fusion strategy is adopted.

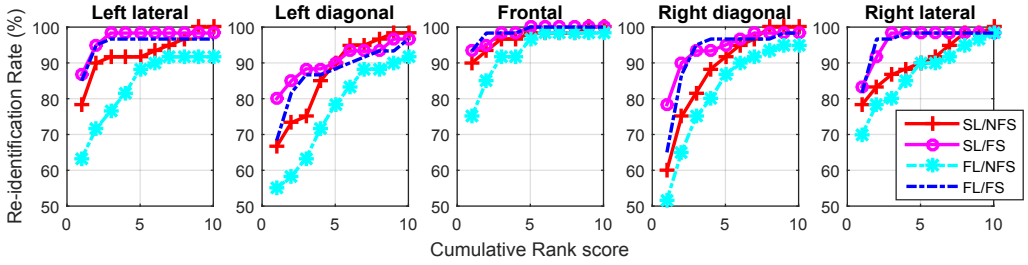


Figure 6.5: The Re-ID performances of various Fusion-FS schemes mentioned in Fig. 6.3 along five contexts *viz.*, left lateral($\sim 0^\circ$), left diagonal($\sim 30^\circ$), frontal ($\sim 90^\circ$), right diagonal($\sim 130^\circ$) and right lateral($\sim 180^\circ$) respectively. Cumulative matching scores up to 10 subjects are shown.

level resulted in lower Re-ID rate (75% illustrated by bold red Dash-dot curve in Fig. 6.4) lying in between the anthropometric and gait fusion results. This under-performance could be ascribable to the large number of potentially misleading irrelevant/redundant features in the feature vector. To tackle this issue, we applied feature selection strategy by exploiting SFS algorithm as explained in Section 6.2.3 and carried out its FS-enabled counterpart Fig. 6.3 (b) Feature Level fusion with Feature Selection (FL/FS).

Various Fusion-Feature selection schemes

After observing the lower performance of the multi-modal system without feature selection, we thereafter carried out an extensive analysis on different fusion-FS schemes as mentioned in Section 6.2.3.Within this set of assessment studies, we carried out all the four fusion-FS schemes *i.e.*, (a) FL/NFS, (b) FL/FS, (c) SL/NFS and (d) Score Level fusion with Feature Selection

(SL/FS), leveraging both feature level/score level fusion and without/with FS. The performance results of all those case studies are illustrated in Fig. 6.5. The corresponding cumulative ranked list (showing anthropometric, gait and overall CMC rank-1) is also shown in Table 6.2. These experimental results corroborate that:

- *Feature Selection (FS) improves Re-ID accuracy, compared to without FS (NFS).*
- *Score-level fusion works better than the feature level fusion in Re-ID.*
- *Overall performance of SL/FS is the best among the group and thus is considered as the 'de-facto' in our context-aware ensemble fusion framework, at the individual classifier bench.*

Table 6.2: Chart showing the Re-ID accuracy rates for five contexts. The accuracy rates shown in each cell represent Anthropometry based Re-ID, gait based Re-ID and Overall Re-ID respectively, at rank-1 CMC. The highest Re-ID rate observed is highlighted in bold letters.

Context	FL/NFS	FL/FS	SL/NFS	SL/FS
Left	**63.33**	61.67	**63.33**	58.33
Lateral	53.33	81.67	53.33	**83.33**
	63.33	85.00	78.33	**86.67**
Left	**75.00**	63.33	**75.00**	63.33
Diagonal	48.33	**61.67**	48.33	**61.67**
	55.00	68.33	71.67	**80.00**
Frontal	**85.00**	78.33	**85.00**	78.33
	58.33	**71.67**	58.33	70.00
	75.00	**93.33**	91.67	**93.33**
Right	**66.67**	**66.67**	**66.67**	**66.67**
Diagonal	46.67	**66.67**	46.67	63.33
	51.67	65.00	66.67	**78.33**
Right	40.00	45.00	40.00	**48.33**
Lateral	63.33	81.67	63.33	**85.00**
	70.00	81.67	80.00	**83.33**

Context-specific FS ensembles

Based on the results obtained, we attribute SL/FS as the de-facto strategy in our framework. The score level fusion of the selected features (both gait and anthropometric) were used to train individual classifiers for each context, at the classifier bench. Also, at this training phase, we also comprehended the relevant features for each context. For analysing the same, we conducted a holistic FS criteria with Cross-validation scheme in each context, which resulted in Table. 6.3. This shows the context-specific features selected for each individual classifier, and using these results, the classifier bench is trained for future evaluation.

It is quite remarkable that the impact of globally discriminative anthropometric features such as height, arm length, chest size are highly relevant in almost all the contexts. However, some features clearly show its influence dependent on the context. For example, gait features presenting angular evolution (hipAngle) and distance showing various right-left limbs during the gait (knee distance, hand distance, elbow distance etc.) were selected in the frontal view. At the same time, many other gait features such as stride length, vertical position evolutions at various joints (headY_μ, kneeY_μ, spineY_μ,hipY_μ, handY_μ etc.) clearly exhibited the evidence of their influence in the lateral contexts. As a consequent result, lateral cases bestowed higher performance of gait features against anthropometric features in our experiments (see SL/FS results in Table 6.2 where LL and RL achieved 83.33% and 85.00% gait based Re-ID

accuracies respectively against corresponding anthropometric based Re-ID accuracies 58.33% and 48.33%), in contrast to the frontal cases where anthropometric features showed better Re-ID performance (Anthropometry based Re-ID is 78.33% against gait based Re-ID of 70.00%). This results clearly corroborate the reason behind why usually gait analysis techniques are better manifested in lateral view rather than in front view.

Table 6.3: Context-specific features selected via SL/FS scheme, during the training of individual context classifiers. Only 28 feature subset out of whole 74 features were selected.

Feature	LL	LD	F	RD	RL	Feature	LL	LD	F	RD	RL
height	✓	✓	✓	✓	✓	spineY_{μ}	✓				
arm	✓	✓	✓	✓		lhipY_{μ}	✓				
upper	✓			✓		lkneeY_{μ}	✓	✓		✓	✓
lower		✓		✓	✓	rkneeY_{μ}	✓	✓		✓	
ULratio		✓		✓		rankleY_{μ}				✓	
chestsize		✓	✓	✓	✓	lhandX_{μ}			✓		
hipsize	✓		✓	✓		lhandY_{μ}	✓				
hipAngle			✓			$\text{lhandY}_{\text{SD}}$				✓	
$\text{kneeDist}_{\mu,\text{SD}}$			✓			rhandY_{μ}				✓	✓
$\text{handDist}_{\mu,\text{SD}}$			✓			lshouldY_{μ}	✓				
elbowDist_{μ}		✓	✓	✓		$\text{lshouldY}_{\text{SD}}$		✓			
$\text{elbowDist}_{\text{SD}}$				✓		rshouldY_{μ}					✓
headY_{μ}	✓	✓	✓		✓	$\text{rshouldY}_{\text{SD}}$			✓		
headY_{SD}			✓			strideLength	✓				✓

6.3.2 Context-Specific Score Level Fusion

After the training of each individual classifier leveraging the context-specific features selected as per in Table 6.3, we conducted testing of our proposed method in pose-invariant scenario where test sample could be at any arbitrary context. In all of our test experiments, we employ a leave-one-out evaluation strategy, where we select one sample at a time and is compared against the rest of the samples in the gallery. This procedure is iterated throughout all the samples in the dataset.

Context detection

When a test sample at an arbitrary context enters into the system, the foremost stage is to detect the context of the test sample by enabling a Context detector module (see Section 6.2.3). Then, based on the detected context, corresponding context-specific classifiers are activated. In order to enable this, a prerequisite was to empirically verify the actual contextual view-points existing in our global dataset, and thus define *'contexts'* based on the gross view-points along a particular direction. So, in order to comprehend the existing contexts in our dataset, a prior context analysis was carried out on our global database, which resulted in context clusters as shown in Fig. 6.6. This empirical analysis enabled us to obtain better insight of the actual view-points spread within each contexts. Based on this study, we could observe that five contexts $\mathbf{v_1}, ..., \mathbf{v_5}$ are spread around their respective clustermeans $\mu = [1.67, 35.63, 92.83, 130.70, 180.17]^\top$, with standard deviations $\sigma = [3.64, 4.90, 3.29, 5.34, 3.99]^\top$.

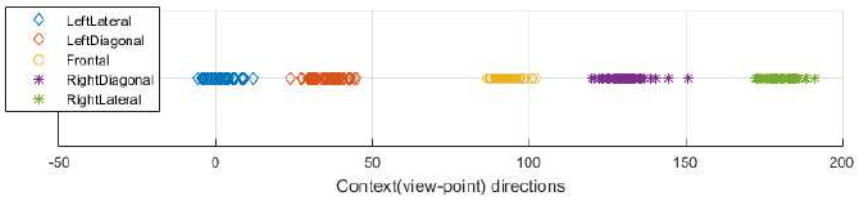


Figure 6.6: Distribution of the context the dataset.

Context-specific fusion strategy

In this fusion stage, information from different context specific classifiers are fused using the runtime estimate of the current context, given by the context detector. After the context detector determines the current context, the corresponding ensemble classifiers are activated. Under this context-aware paradigm, two schemes were proposed: (i) Using ***single context (binary weighted)***, in which only the closest context is selected based on the nearest cluster mean among all the clusters. Hence, only that specific context in the gallery is activated and the test is matched against all the rest of the 59 samples in that particular context and the ranked Re-ID list is obtained. The second context-aware scheme is (ii) using ***Two contexts (linear interpolated weights)***, wherein depending upon the probe context, two nearby contexts (between which the test probe lies) are activated. Then, the test sample is matched against the rest of the gallery samples in those two contexts, and respective matching scores are generated[6]. Then, depending on the distance of the test context with respect to those contexts, adaptive weights are assigned via linear interpolation technique, and Context-specific score level fusion strategy is applied to obtain the aggregate Re-ID result (see Section 6.2.3).

In order to perform a comparison of our context-aware proposal, we also conducted baseline studies without the notion of context (Context-unaware). In these baseline scenarios, we disabled the context detector module, and hence no notion of the probe context is available to the system. Three baseline studies were performed. In the first case study, the test sample enters into the system and it matches against all the rest of the 299 gallery samples (from all the contexts), and computes the ranked Re-ID list based on the matching score. In this method, albeit the testing is performed as context-unaware, the features used per person are context-aware. In other words, the samples per person used in the test mode were trained a priori based on the context-specific feature selection. Hence, we term this scenario as ***'Pseudo baseline'***. To tackle this context dependency, we conducted the second case study called ***'Pure baseline'***, where we made the system context-unaware not only at the testing phase, but also at the training phase. In order to conduct this analysis, we retrained our system and applied global feature selection upon all the samples, independent of the context. Thus, the same features got selected globally, thus making the FS in all the samples context-unaware. Afterwards, testing was conducted as in the *'Pseudo baseline'* case, where the probe is matched against all the rest of the 299 gallery samples, and computes the ranked Re-ID list based on the matching score. The third context-unaware case study is with the assumption that the

[6]Since the number of gallery samples per context may vary (the actual context containing text context contains 59 samples in the gallery whereas the other context contains 60 samples), the matching scores per sample are of different size. Hence, we use matching scores per person by computing the best score (minimum distance score) per person.)

Table 6.4: Results of classifier fusion showing our proposed context-aware classifier fusion against context-unaware baseline case studies. In context-aware cases, context detector module is enabled, whereas in the context-unaware cases, context detector module is disabled

	Context-unaware			**Context-aware**	
	No context (Pseudo baseline)	**No context (Pure baseline)**	**All contexts (equal weights)**	**1 context (binary weights)**	**2 contexts (adaptive weights)**
Anthropometric	25.33%	60.33%	45.67%	68.67%	68.00%
Gait Re-ID	26.67%	70.33%	53.33%	84.67%	85.67%
Overall Re-ID	74.33%	79.33%	71.33%	88.67%	88.33%
Processing time	25.7sec.	21.64sec.	25.92sec.	5.59sec.	10.47sec.

chance of the probe sample within the pre-defined contexts are equally likely *i.e.*, the same probability of occurring. Hence, ***equal weights*** of 0.2 is assigned to each contexts. The probe sample is tested against the gallery samples in each context and then weighted sum upon all the five individual classifier matching scores are performed to obtain the aggregate matching score and the consequent ranked list.

The results of all the five case studies mentioned above are shown in Table 6.4. It is quite remarkable to observe that, context-aware methods (either by using a single or two contexts) bestow high performance level ∼88%, whereas all variants of Context-unaware cases miss good results ∼71%-79%. Also, since there is no notion of the context in Context-unaware cases, the probe sample has to be matched against all the rest 299 samples in the global dataset. At the same time, in context-aware cases, the information of the direction helps to reduce the size of the gallery set drastically by making it context-specific. Due to this reason, context-aware systems performed faster (∼5-10 sec.) compared to the context-unaware system (∼21-25 sec.). This highly accentuates the fact that, in unconstrained scenarios, the knowledge of context can augment the performance of a Re-ID system in terms of both speed and accuracy.

6.4 Summary

In this work, a novel context-aware ensemble fusion framework has been proposed towards long term Re-ID in the wild. In order to develop this framework, we first analysed the individual as well as fused Re-ID results leveraging anthropometric and gait features. Based on the observation that albeit multimodal fusion improves the result, naïve integration of large number of potentially irrelevant features can cause degradation of results, we proposed a Feature selection (FS) technique by employing Sequential Feature selection (SFS) algorithm. In this regard, various fusion-FS strategies were analysed and the best among all (SL/FS) has been selected as the *de facto* standard in our framework.

Another contribution was the concept of context-specific classifiers. This was quite significant depending upon the property of the sensor that, specific features are well acquired in specific directions. Based on our FS scheme, we adaptively selected those features depending upon the directions which we term as *'contexts'*, and trained each individual classifiers based on the selected features for that particular context. During the run time, the direction of the probe sample was determined using a Context-detector module, and the corresponding neighboring context/contexts were activated. Afterwards, a context-aware classifier fusion was facilitated via our proposed *'Context-specific score level fusion'*, and the Re-ID was carried out. The

experimental results showed that comparing to the Context-unaware systems, context-aware systems performed significantly faster (up to 4.5 times) and accurate (up to 17 percentage point better).

In the future works, we envisage to extrapolate this study by collecting more data in more random directions of walk (moving from a denser context clusters to scatter clusters), and to analyse how the linear interpolation strategy can enhance the results. Another idea is also to incorporate multiple contexts in the scenario, (*i.e.*, in addition to the view-point, also include distance to the camera, occurrence of face, person co-occurances etc.) in order to improve the re-identification performance.

Chapter 7

Conclusions and Perspectives

I think and think for months and years. Ninety-nine times, the conclusion is false. The hundredth time I am right.

— Albert Einstein

This thesis proposes a few step changes in the problem of long-term and view-invariant person Re-Identification (Re-ID) by exploiting soft-biometrics (human anthropometry/shape and human gait), 3D data and contextual methods. We developed two datasets (HDA and KS20 Vislab Multi-View Kinect Skeleton datasets) in realistic scenarios and one synthetic dataset leveraging virtual reality avatars. (refer Section 1.5)

7.1 Key contributions

- **Anthropometry based Person Re-ID:** Our first direct contribution is the proposal of a novel Re-ID framework, for person re-identification either by multimedia data or by means of manual queries describing natural human compliant labels (soft biometric traits). The proposal of such an automatic dual mode system by incorporating 'human in the loop' is quite appropriate in a practical perspective where the operator can opt collecting either multimedia info from the camera or eye witness description of the person to carry out person identification. By exploiting shape context (SC) descriptor extracted on the head-to-torso region on frontal human silhouettes (less occluded and more stable features), its applicability was experimentally confirmed in both real and virtual reality data sets.

  We studied the relationship between Shape Context descriptors and soft biometrics. By means of regression methods, the semantic gap between the manual and machine interpretation of human profile has been analysed and corroborated as 'linear' in nature. The use of digital graphics/animation platform (*Unity3D*®) along with computer vision and machine learning techniques has enabled Re-ID without the need of time-consuming manually annotated data.

- **Gait based Person Re-ID:**

  A flow-based gait period estimation and a novel Histogram of Optic flow Energy Image (**HOFEI**) over the entire body were proposed in this work. The main advantage of such a methodology is that it facilitates online Re-ID, since optic flow based methods do not require silhouette segmentation. No state-of-the-art works addressed Re-ID with optic flow features in frontal gait. The proposed algorithm has been tested on a controlled benchmarking gait dataset (CASIA dataset) and a more challenging benchmarking video surveillance dataset (HDA Person dataset). Various case studies confirmed the effectiveness of the proposed gait based Re-ID technique under challenging conditions of different background clutter and sampling rates (25Hz in CASIA vs 5Hz in HDA).

- **Towards multi-modal and view-point invariant Person Re-ID:** We have studied the effect of the combination of multiple features in Re-ID. We used both static and dynamic features related to the human posture and walking, in order to select which features are better to classify people in the population. It was observed that the static anthropometric features are more informative than gait features when employed individually. However, when fusing many static anthropometric features and dynamic gait features, the overall recognition accuracy increases.

  A key contribution was the first actual demonstration of the of the real impact of view-point on the Re-ID paradigm using data acquired from the Kinect sensor. Traditional studies assume the walking sequences to be in a specific view-point. Instead, we conducted a view point invariant benchmark assessment by experimenting explicitly different view-points in the probe and gallery samples. Various case studies conjectured that the view-point has a great influence on the feature extraction and the Re-ID performance will vary according to the selection of view-points in the gallery and probe sets.

- **Context-Aware Person Re-ID:** A novel context-aware ensemble fusion framework for long term Re-ID has been proposed. Realizing that the computation of biometric features depend strongly on the view-point, we incorporated the information associated to the view-points termed as 'contexts' and proposed the so called 'Context-aware ensemble fusion Re-ID framework'.

  Assigning view-points to contexts, feature selection technique (SFS) can to fine tune the most relevant features in each context that are used to train context-specific classifiers. Various feature selection and fusion strategies have been a part of the research in order to understand the best combination of features in each context. The observed superior performance of context-aware systems both in terms of speed and accuracy, look quite promising.

7.2 Limitations and Future works

There are still many open issues associated with the Re-ID problem, which we couldn't explore completely during this thesis period. However, we envisage to include some in our future works.

– Albeit we addressed the long term re-identification problem, by leveraging soft-biometric cues, the possibility of exploring many other cues are still open. For instance, **face information** could improve the Re-ID rate drastically. Even though face information can't be exploited in all the viewing directions (for example, rear/lateral cases), some context enabled fusion of face and other biometric cues could be envisaged. In the same way, the idea of using human shape in the Re-ID frame work could be extended towards **full body shape**. This would provide a better representation of the 3D shape of persons, and lead to more accurate re-identifications.

– Another possibility is to leverage some tecniques to understand the dressing style and the choices of apparels. For example, personal choices of the dress colours, signature apparels, ornaments/other accessories can also contribute towards the process of distinguishing among the population. Some celebrities' dress code are shown in Fig.7.1. In future, deep learning techniques could be able to comprehend these kind of abstract cues to be incorporated with the process of re-identification. Due to the availability of increasing amount of big data and deep techniques, such abstract level information can be deployed to the Re-ID paradigms.

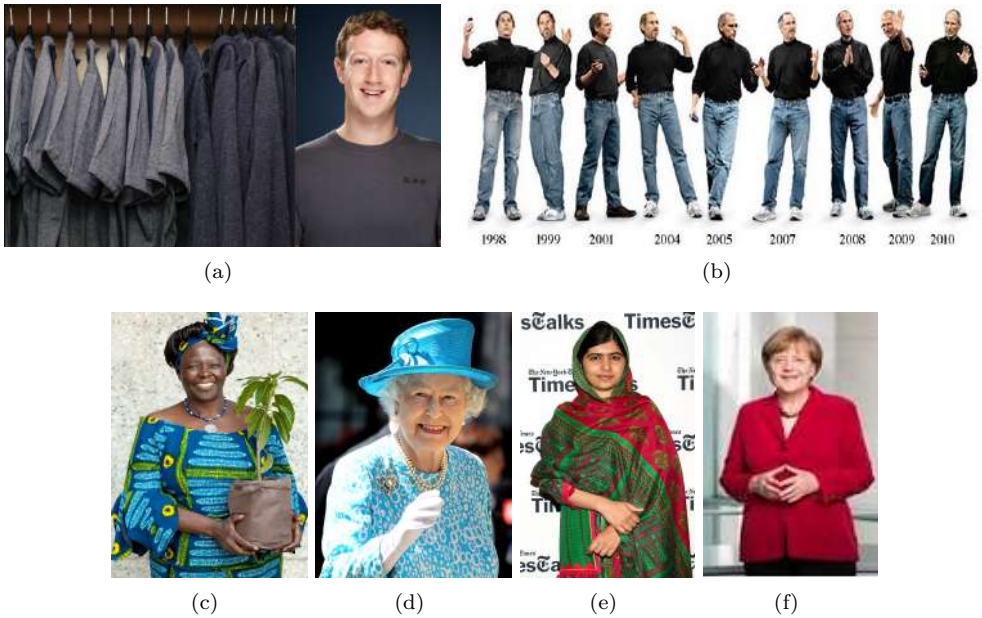


Figure 7.1: Different dressing attires also could be exploited towards Re-ID process via deep learning techniques. The figure above shows the typical dress codes of some famous people around the globe (a) Mark zuckerberg (b) Steve Jobs (c) Wangari Maathai (d) Queen Elizabeth (e) Malala Yousafzai and (f) Angela Merkel. Rather than just the colour/texture of the appearance, the style of appearance also matters, and more interestingly also could be utilized towards long term Re-ID.

–On gait based Re-ID, there are many open issues. Due to the wild nature of the re-identification scenarios, gait based Re-ID would benefit a lot from true pose invariant approaches to gait analysis. Even though some pose invariant approaches have been proposed using 2D images [Wang *et al.* , 2016; Wei *et al.* , 2015] or 3D models generated out of mul-

tiple 2D cameras [Iwashita *et al.* , 2010], their levels of accuracy still lag behind what can be achieved with MOCAP data, therefore leaving room for improvements. Recently, many state-of-the-art methods address the pose-invariant 3D data generated by Kinect-like devices [Chattopadhyay *et al.* , 2015; Andersson & Araujo, 2015], including our proposal [Nambiar *et al.* , 2017b]. These latter techniques have been quite revolutionary in terms of data acquisition, as well as the classification accuracy, however, it is not clear how to exploit them in traditional surveillance scenarios which use 2D cameras. Hence, one possible future direction is to research further on **reconstructing 3D data from the traditional 2D image frames**.

– Another significant problem is the **open-set Re-ID**. In the open-set scenarios the system should be able to detect novel subjects, i.e. persons not yet enrolled in the gallery [Gala & Shah, 2014]. Only a few works on person Re-ID address the open space scenario [Liao *et al.* , 2014; Bedagkar-Gala & Shah, 2011]. Similarly, although gait is well suited for long term person identification, only a few works have verified their performance over the longer periods. All of them suffer degradation in the Re-ID performance with the changes in the covariates (*e.g.*, different terrains, wear accessories, seasonal variations in the dressing styles) compared to their performance over the short period. This problem motivates **the necessity for more long-term gait based Re-ID datasets**, where algorithms could be tested against seasonal variations, and more real-world experiments in long-term scenarios.

–Regarding the Context-aware Re-ID, we envisage to **incorporate multiple contexts in the scenario**, *i.e.*, in addition to the view-point. By including other contexts such as camera topology, distance to the camera, types of environments under surveillance, types of activities under consideration (leisure, work, passage), the amount of interactions among persons etc., we expect to enrich the the quality of the methods. An extension of the current work is also in preparation, to exploit information across contexts (cross-contextual analysis). In many cases the probe data appears in contexts where the gallery samples of the person are few. So we should gather information for the identity of a person in multiple contexts. Our approach is to study which features better map among different contexts. The publication will be submitted soon. [Nambiar *et al.* , 2017 **(In preparation)**]

Appendix A

HDA Person dataset

A.1 HDA Person dataset

High Definition Analytics (HDA) dataset was designed with the following goals: (i) establishing a benchmark for Pedestrian Detection (PD) algorithms specific for an office scenario, (ii) providing a benchmark featuring High Resolution images for Video Surveillance algorithms, in particular PD, person tracking and Re-Identification (Re-ID), and (iii), creating a benchmark for fully automated Re-Identification (PD+REID) systems. We think that the availability of a benchmark for PD algorithms in an office scenario will attract the attention of the Video Surveillance community on PD's. The use of cameras equipped with both standard and High Definition sensors will permit the study of the effect of High Definition on the performance of the algorithms. Moreover, the presence of Hight Resolution images will highlight the weaknesses of the Video Surveillance algorithms of the current generation for that specific case and foster the development of algorithms specific for High Definition images. To make a concrete example, we expect that the algorithms in the state of the art not to achieve real time performance on High Resolution images. Finally, we think that the creation of a benchmark for PD+REID will help the establish a community for the study of this problem, which we see as the natural evolution of classic Re-ID.

The HDA dataset was acquired recording simultaneously from 13 indoor cameras for 30 minutes. The cameras were distributed over three floors of the Institute for Systems and Robotics (part of the Instituto Superior Técnico in Lisbon, Portugal), a typical office scenario for Video Surveillance. Approximately 85 people participated in the data collection, most of them appearing in more than one camera. The dataset is heterogeneous: we used three distinct types of cameras (standard, high and very high resolution), different view types (corridors, doors, open spaces) and different frame rates. This diversity is essential for a proper assessment of the robustness of video analytics algorithms in different imaging conditions.

The data recordings for the HDA dataset involved the use of 13 AXIS cameras, some with standard VGA resolution (AXIS 211, AXIS 212PTZ, and AXIS 215PTZ) some with 1MPixel resolution (AXIS P1344) and one of 4MPixel resolution (AXIS P1347). To save bandwidth, storage and labelling time, the sequences were not acquired at high frame rates, but at rates of 5Hz, 2Hz and 1Hz for the VGA, the 1MPixel and the 4MPixel resolution respectively. The camera poses in the three floors are depicted in Figure A.1. Table A.1 describes the camera

network details in brief. Figure A.2 displays one frame for each camera, highlighting differences in illumination, color balance, depth range and camera perspective.

(a) Floor 6 (b) Floor 7

(c) Floor 8

Figure A.1: Camera poses: a visualization of the three floors of the building at which the HDA dataset was acquired. The cameras marked with a red circle and an orange field of view are the ones used to record data.

Table A.1: Details of the labelled camera network.

CAM	02	17	18	19	40	50	53	54	55	56	57	58	59	60
640x480	✓	✓	✓	✓	✓									
1280x800						✓	✓	✓	✓	✓	✓	✓	✓	
2560x1600														✓
fps	5	5	5	5	5	2	2	2	2	2	2	2	2	1
floor	6	8	8	8	8	7	7	7	7	7	7	7	8	7

A.2 Labelling for the HDA dataset

The labelling for the HDA dataset consists in Bounding Boxes (BB's) associated with a unique person identifier (ID) and an occlusion flag. Each person/group of people in the images is labelled by such a BB. We opted for using an occlusion flag instead of a value encoding the occlusion ratio of a person because of the much faster annotation process required by the former: given the elevated number of annotations in the dataset, this choice made the labelling

(a) Camera 02

(b) Camera 17

(c) Camera 18

(d) Camera 19

(e) Camera 40

(f) Camera 50

(g) Camera 53

(h) Camera 54

(i) Camera 56

(j) Camera 57

(k) Camera 58

(l) Camera 59

(m) Camera 60

Figure A.2: Snapshots of the sequences acquired in the HDA dataset. Notice the differences in illumination, color balance, depth range and camera perspective.

task more manageable. The BB's alone are used as Ground Truth (GT) in the PD task, while the information conveyed by the BB's needs to be augmented by the person ID for evaluating the Re-ID algorithms. The GT for benchmarking tracking algorithms is encoded by the ID of the BB's, together with the initial and final frame for each person appearance in a video sequence. In the process of labelling, we used the following software tools: MATLAB® with the Image Processing Toolbox, Piotr Dollár's Toolbox [Dollár, n.d.b] and Detection Code [Dollár, n.d.a].

This is the list of the labelling rules:

1. Each BB is drawn so that it completely and tightly encloses the person.
2. If a person is partially occluded, the BB is drawn estimating the whole body extent.
3. Truncated people (i.e., people with projections partially outside the image boundaries) have their BB's cropped to image limits.
4. The occlusion flag is set to '0' for fully visible people, while for partially occluded and truncated people it is set to '1'.
5. A unique ID is associated with each person. In case determining the identity of a person is impossible for the labeller, the special ID 'personUnk' is used.
6. Groups of people that are impossible to label individually are labelled collectively as 'crowd'. People in front of a 'crowd' area are labelled normally.

The proposed labelling allows researchers to perform different experiments on a single test set. For instance, one could choose to test one algorithm ignoring Missed Detections on heavily occluded people, or detections on crowded regions.

We show examples of labelling in Figure A.3. The person ID is indicated at the top of each BB. The HDA dataset comprises annotations of 85 persons, of which 70 are men and 15 are women. A statistical characterization of the data is presented in Table A.2 and Figure A.4. One of the peculiarities of the HDA dataset resides in the exceptionally wide range of peoples' BB heights: from 69 to 1075 pixels (see Figure A.4(c)).

(a) (b) (c)

Figure A.3: Labelling examples. (a) A fully visible (unoccluded) person. (b) Two partially occluded people. (c) A crowd with three partially occluded people in front of it. The ID of each person is indicated on top of the Bounding Boxes.

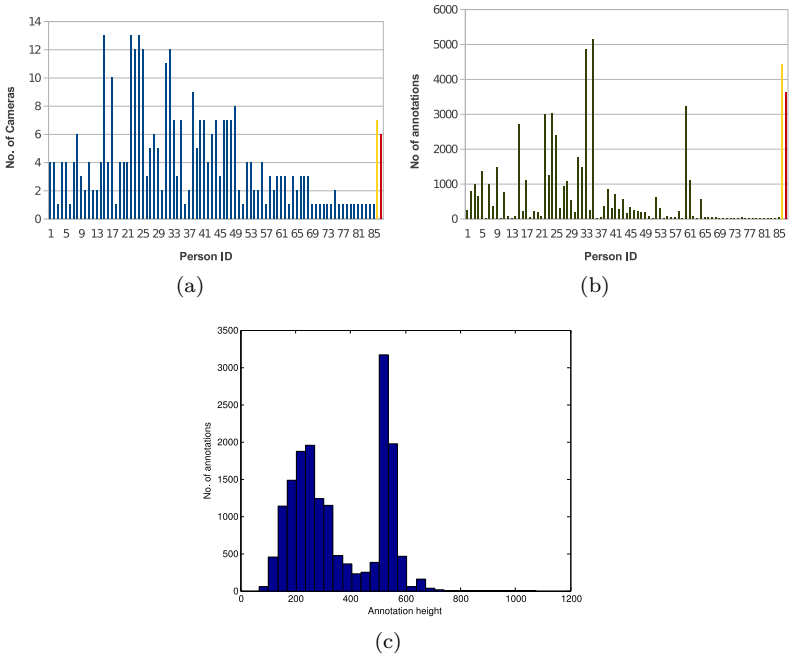


Figure A.4: (a) Number of sequences each person appears in. Person 86 (yellow) and 87 (red) correspond to the labels 'personUnk' and 'crowd'. (b) Number of Bounding Boxes (BB's) for each person. (c) Histogram of BB height for the unoccluded people. The peaks of the VGA and the high resolution distributions are visible. The BB's span heights between 69 and 1075 pixels.

Table A.2: Data on the number of frames, the number of annotations and the number of people for each sequence. The minimum and maximum height of unoccluded Bounding Boxes (BB's) are also reported. Camera 02 does not have person height information due to its unconventional overhead perspective.

Camera	02	17	18	19	40	50	
# frames	9819	9897	9883	9878	9861	2227	
# BB's	1832	3865	13113	18775	7855	1288	
Min. height	-	310	90	71	71	158	
Max. height	-	463	338	403	408	606	
# persons	9	26	32	34	39	20	
# frames	3521	3424	3798	3780	3721	3670	1728
# BB's	465	8703	576	3190	2291	894	1182
Min. height	69	153	619	384	395	598	212
Max. height	681	608	717	688	681	775	1075
# persons	19	12	34	43	34	34	20

A.3 Access to the data

The link for HDA dataset is http://vislab.isr.ist.utl.pt/hda-dataset/. Access to the dataset is available upon request. We received 74 requests so far.

Regarding the data format, for each camera we provide the .jpg frames sequentially numbered and a .txt file containing the annotations according to the "video bounding box" (vbb) format defined in the Caltech Pedestrian Detection Database[1]. Also on this site there are tools to visualise the annotations overlapped on the image frames.

Some statistics info are also mentioning herein:

Labeled Sequences	13
Number of Frames	75207
Number of Bounding Boxes	64028
Number of Persons	85

[1] Caltech Pedestrian Detection Database: http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

Appendix B

Kinect based Re-ID dataset

B.1 Our dataset: KS20 Vislab Multi-view Kinect Skeleton dataset

In order to employ Re-ID in a realistic 'in-the-wild' scenario, it is quite essential to have a challenging unconstrained dataset, comprised of sequences of people walking in different directions. Since such a Kinect based dataset (with different view angles) towards gait based Re-ID was unavailable, we acquired our own dataset using a mobile platform, in the host laboratory. It is made publicly available now, in the host webpage http://vislab.isr.ist.utl.pt/vislab_multiview_ks20/

For the data acquisition, we used a mobile platform, in which the kinect sensor was fixed at a height of an average human. (See Fig. B.1(a) for the data acquisition system)

Figure B.1: Data acquisition system set up

Originally, the acquisition set up has been developed as a part of AHA(Augmented Human Assistance) project (http://aha.isr.tecnico.ulisboa.pt/), in our host laboratory. The AHA Recording System was designed to acquire and record data from the movement and body signals of humans performing exercises according to the AHA Exercise Protocol. It fulfills two objectives in the AHA project: to record exercise data for offline analysis and

algorithm development and; to provide real time data streaming for online analysis in the robotic platform.

In addition to the KINECT sensor, it consisted of 2 omnidirectional cameras, Bitalino and Bioplux wearable biosensors, a PC with Bluetooth dongle (to acquire from Bitalino or Bioplux modules), LCD Display and wireless keyboard and 8 ports switch to allow the usage of multiple ip network cameras. All the above sensors were acquired via YARP drivers that stream time stamped information to ports. Currently, we used only the Kinect sensor among them and the PC to record the data streams. And the acquired data streams from KINECT are the following:

- ahaKinectColor: Images from the Kinect camera. RGB 8 bit unsigned integer.
- ahaKinectDepth: Depth data from the Kinect sensor. Single channel 8 bit unsigned integer.
- ahaKinectBody: Skeleton coordinates arranged according to the file Yarp_Messages_Structure described below in Section B.1.1.

B.1.1 Yarp Messages Structure for Kinect v2 Body Frame:

The Kinect v2 body frame data is streamed through the buffered port "/unityServer/kinectv2/body:o" at 30Hz. The data is encapsulated in a bottle with the structure defined as follows.

- **Bodies Bottle Structure**: The Bottle encapsulating all the data from Kinect v2 Body frame. It holds multiple bottles (0 to 6 bottles) each encapsulating the data for one tracked body. The maximum number of bodies tracked by the Kinect v2 is 6.

  ((Body_1) (Body_2) ... (Body_N))

  Bottle Body_X – Bottle that holds the data of one tracked body.

- **Body Bottle Structure**: Holds the data of one tracked body, includes the Tracking Id of the body and 25 Joint Bottles describing the Kinect v2 skeleton.

  (Tracking_Id (Joint_1) (Joint_2) ... (Joint_25))

  String Tracking_Id – Unsigned long integer identifying the body being tracked converted into string type.

  Bottle Joint_X – Bottle that holds the joint data, in Kinect v2 each body has 25 joints.

- **Joint Bottle Structure**: Bottle holding the data of a joint, includes joint type, tracking state, joint position and joint orientation.

  ((Type) (Tracking) (Position) (Orientation))

  Bottle Type – Bottle storing the joint type data.

  Bottle Tracking – Bottle storing the joint tracking data.

  Bottle Position – Bottle storing the joint position data.

  Bottle Orientation – Bottle storing the joint orientation data

- **Type Bottle Structure**: Bottle storing the joint type data both in integer and string format.

  (Type_Int Type_Name)

  Integer Type_Int – Integer identifying the joint type.

  String Type_Name – String identifying the joint type.

  The Kinect v2 Joint Type follows the enumeration:

Type_Name	Type_Int	Type_Name	Type_Int	Type_Name	Type_Int
SpineBase	0	ElbowRight	9	KneeRight	17
SpineMid	1	WristRight	10	AnkleRight	18
Neck	2	HandRight	11	FootRight	19
Head	3	HipLeft	12	SpineShoulder	20
ShoulderLeft	4	KneeLeft	13	HandTipLeft	21
ElbowLeft	5	AnkleLeft	14	ThumbLeft	22
WristLeft	6	FootLeft	15	HandTipRight	23
HandLeft	7	HipRight	16	ThumbRight	24
ShoulderRight	8				

Figure B.2: Kinect v2 Joint Type enumeration

- **Tracking Bottle Structure**: Bottle of a single element describing the tracking state of a joint.

  (Tracking_State)

  Integer Tracking_State – Integer describing the joint tracking state.

  The Kinect v2 Joint State follows the enumeration:

Tracking_State	Enum
NotTracked	0
Inferred	1
Tracked	2

Figure B.3: Kinect v2 Joint Type enumeration

- **Position Bottle Structure**: Bottle storing the Cartesian coordinates of a joint. The Kinect v2 reference frame is as indicated in Figure B.4, where X points to the left of device, Y upwards and Z forward.

  (X Y Z)

  Float X – X coordinate of the joint in meters.

  Float Y – Y coordinate of the joint in meters.

  Float Z – Z coordinate of the joint in meters, distance to the X0Y sensor plane.

- **Orientation Bottle Structure**: Bottle storing the quaternion values that define the joint parent bone orientation relative to the Kinect v2 reference frame, Figure B.4. Each

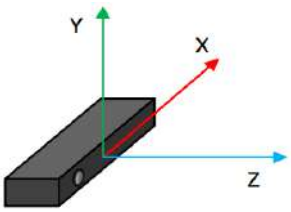


Figure B.4: Kinect V2 camera space reference frame

bone reference frame, depicted in Figure 2, is defined as follows: Origin located at the bone's child joint; Y axis (bone direction) collinear with the bone, directed from the previous joint towards the current joint; Z axis (normal) perpendicular to the bone, collinear with the joint roll axis; and X axis (binormal) perpendicular to Y and Z, forming a right handed reference frame.

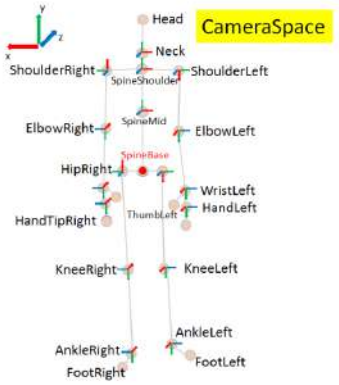


Figure B.5: Kinect V2 body joints reference frames

Example: the orientation of the bone connecting the right hip to the right knee is given by the "KneeRight" joint reference frame orientation relative to the camera space reference frame. **(X Y Z W)**

Float X – X component of the Quaternion

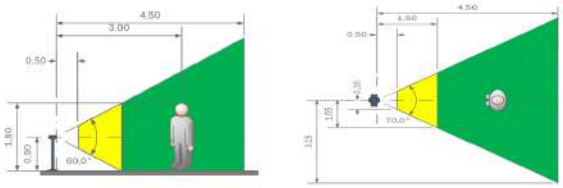
Float Y – Y component of the Quaternion

Float Z – Z component of the Quaternion

Float W – W component of the Quaternion

B.1.2 Sensor Placement:

The sensor will be placed facing the user at about 3m from it, at a height of about 0.9m as shown in the following figures where the green area represents the space where full body tracking is possible.



(a) (b)

Figure B.6: Kinect v.2 placement: (a) Front view (b) Top view

B.2 Instructions to perform recordings

- **Begin an acquisition**

  1. Connect the system to a power plug and press the button on the power strip.
  2. Turn on the PC.
  3. Launch the AHA acquisition module double clicking the light blue icon on the desktop.

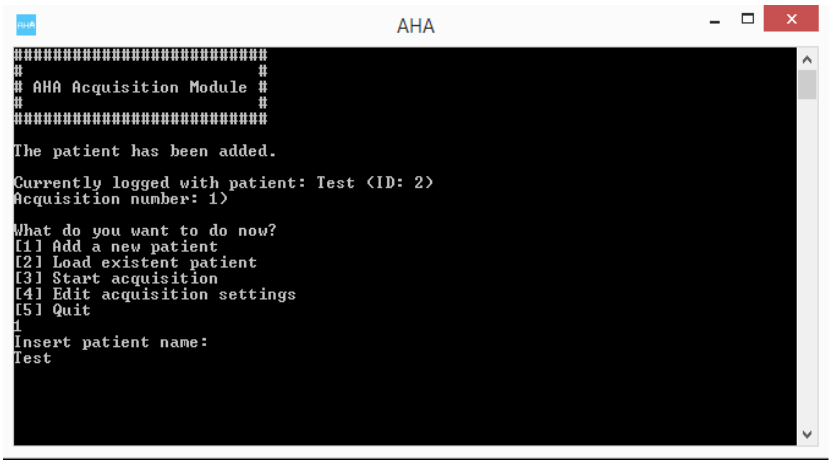
  4. Wait until the system finishes the loading procedure.

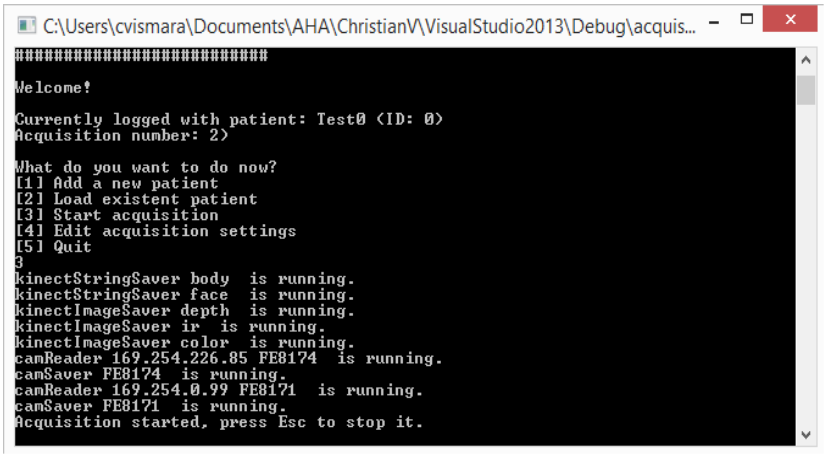
  5. Press “1” and “Enter” to add a new patient
  6. Insert patient name and press “Enter”
  7. Press “3” and “Enter” to begin the acquisition
  8. Press “Esc” when you want to end the acquisitions

- **Change Patient**

  Add a new patient:

  1. From the main menu, press “1” and “Enter”

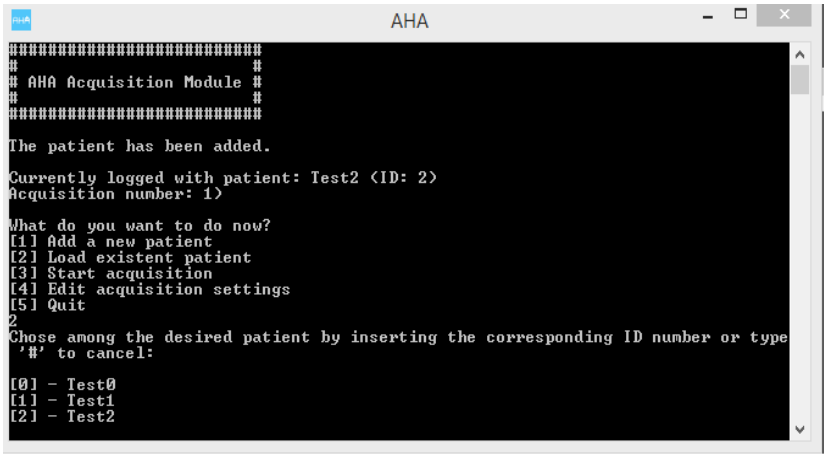




2. Insert patient name and press “Enter”

Load an existent patient:

1. When you are in the main menu, press “2” and “Enter”
2. Choose the desired patient ID and press “Enter”



- **Change acquisition settings**

1. From the main menu, press “4” and “Enter”
2. Press the number corresponding to the setting you want to edit “1” is to adjust camera acquisition frequency (default:1Hz) “2” is to adjust string acquisition frequency (default: unlimited), “3” is to choose the device to acquire biosignals (default: both off), “4” is to choose the omnidirectional camera (default: both on)
3. Insert the new value and press “Enter”
4. Press “0” and “Enter” to go back to the main menu.

- **Quit the application and turn off the system**
  1. From the main menu, press “5” and “Enter”
  2. Turn off the PC
  3. Press the button on the power strip and disconnect the system from the power plug

B.3 Instructions to verify recordings

The software will check during data acquisition that everything is properly working but still, there are several ways to check that the system is correctly performing the data acquisition:

- Check using mongo client if any database is missing (command “show dbs”). Check the appendix on further instructions about mongodb.

- Check using mongo client that the number of elements in every collection is growing (command “use $< database - name >$” and then “db.$< collection - name > .count()$”)

- Visualize data at the end of the data acquisition using matlab scripts (check next section)

B.4 Instructions to use recordings on a different computer

After data is recorded, it can be retrieved for offline use. This section describes the instructions to prepare a different computer to use the acquired recordings.

-Download MongoDB from here http://www.mongodb.org/downloads

-Create a /data/db folder in some location in your disk.

-In a terminal window go to the folder where you have the mongodb binaries

-Run ”$mongod - -dbpath < path - to - data - folder >$ ” from terminal.

-Run ”$mongorestore - -db < db - name >< path - to - dbfolder >$ ” from terminal, e.g. mongorestore –db ahaKinectBody

-(Optional) Run ”mongo” from terminal and then execute the command ”show dbs” to check if the requested databases have been correctly imported

- Run dataVisualization.m. Before running you may want to comment the lines that visualize the data you’re not interested in

- (Optional) To visualize timestamps set the verbose = 1.

Appendix C

Publications & other scientific activities

Refereed Journals:

- *Athira Nambiar, Alexandre Bernardino and Jacinto C. Nascimento, "Gait based Person Re-identification: a Survey", acm Computing Surveys, 2017 (submitted).*

- *Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento, Ana Fred and Jose Santos-Victor,"A context-aware method towards view-point invariance in 'in-the-wild' long-term Re-identification", 2017. (submitted).*

- *Athira Nambiar, Alexandre Bernardino and Jacinto Nascimento, "Shape context for soft biometrics in person re-identification and database retrieval", Pattern Recognition Letters, 2015.*

- *Athira Nambiar, Matteo Taiana, Dario Figueira, Jacinto Nascimento and Alexandre Bernardino, "A Multi-camera video dataset for research on High-Definition surveillance", International Journal of Machine Intelligence and Sensory Signal Processing, Special Issue on Signal Processing for Visual Surveillance, Inderscience Journal, 2014.*

Refereed Conferences and Workshops:

- *Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento and Ana Fred, "Context-Aware Person Re-identification in the Wild via fusion of Gait and Anthropometric features", 2nd International Workshop on Biometrics in the Wild (BWild), in conjunction with IEEE Conference on Automatic Face and Gesture Recognition, Washington DC, USA, 2017.*

- *Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento and Ana Fred, "Context-aware Person Re-identification via Fusion of Anthropometric and Gait Features", One day BMVA Technical Meetings- Security and Surveillance, British Computer Society,*

London, UK, 2017.

- *Athira Nambiar, Alexandre Bernardino, Jacinto C. Nascimento and Ana Fred, "Towards view-point invariant Person Re-identification via fusion of Anthropometric and Gait Features from Kinect measurements", VISAPP, 12th International Conference on Computer vision Theory and Applications, Porto, Portugal, 2017.*

- *Athira Nambiar, Alexandre Bernardino and Jacinto Nascimento, "Person Re-identification based on Human query on Soft Biometrics using SVM Regression", VISAPP -11th International Conference on Computer vision Theory and Applications, Rome, Italy, 2016.*

- *Athira Nambiar, Jacinto C. Nascimento Alexandre Bernardino, and Jose Santos-Victor, "Person Re-identification in frontal gait sequences via Histogram of Optic flow Energy Image", Advanced Concepts for Intelligent Vision Systems ACIVS, 2016.*

- *Matteo Taiana, Dario Figueira, Athira Nambiar, Jacinto Nascimento and Alexandre Bernardino, "Towards Fully Automated Person Re-Identification ", VISAPP 2014, 9th International Conference on Computer vision Theory and Applications, Lisbon, Portugal, January, 2014.*

- *Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento and Alexandre Bernardino, "The HDA+ dataset for research on fully automated re-identification systems", Proc. of ECCV2014 Workshop on Visual Surveillance and Re-identification, Zurich, Switzerland, 2014.*

- *Athira Nambiar, Paulo Lobato Correia and Luis Ducla Soares, "Frontal Gait Recognition Combining 2D and 3D Data", Proc ACM Workshop on Multimedia and Security - MM-Sec, Conventry, United Kingdom, 2012.*

- *Athira Nambiar, Marco Tagliasacchi and Enrico Magli, "Secure image databases through distributed source coding of SIFT descriptors", IEEE International Workshop on Multimedia Signal Processing (MMSP), Banff, Canada, 2012.*

- *Athira.M.Nambiar, Asha Vijayan and Aishwarya Nandakumar, "Wireless Intrusion Detection Based on Different Clustering Approaches", First Conference Of Women in Computing in India, published in acm portal digital library, India, 2010.*

Projects:

- AHA- Augmented Human Assistance (http://aha.isr.tecnico.ulisboa.pt/) -*Sep 2014- now*
  Supervisor: Prof. Alexandre Bernardino, Vislab, ISR/IST.
- HDA- High Definition Analytics (http://vislab.isr.ist.utl.pt/hda-dataset/) -*Jan 2013-March 2014*
  Supervisor: Prof. Alexandre Bernardino, Vislab, ISR/IST.

- BIOSEC- Secure Multimodal Biometric Recognition System (Human gait analysis) -*Sep 2011- March 2012*
  Supervisor: Prof. Paulo Lobato Correia, Multimedia Signal Processing Group, IT/IST.
- Distributed Source Coding, Database Security -*Sep 2010- June 2011*
  Supervisor: Prof. Enrico Magli, Image Processing group, POLITO.

Other scientific activities:

- Session chair for VISAPP 2017, 12th International Conference on Computer vision Theory and Applications, Porto, Portugal.
- Reviewer for Pattern Recognition Letters, International Joint Conference on Artificial Intelligence (IJCAI).
- Oral conference presentations in VISAPP2016 (Rome, Italy), ACIVS2016 (Leece, Italy), VISAPP2017 (Porto, Portugal), BMVAmeeting2017 (London,UK) and BWild,IEEE FG (Washington DC, US).
- Poster presentations at Annual Meeting of LarSyS (Laboratório de Robótica e Sistemas em Engenharia e Ciência) 2014, 2015, 2016 Lisbon and PhD open days at IST, 2017.
- Presented the work of Prof. Shishir. K. Shah's group, Computer Science Department, University of Houston in VISAPP 2017 conference, on Human activity recognition.
- Project developer (HDA, AHA) and dataset managing & distribution team member
- Robot exhibition assistance in Ciencia2016 Portugal, RoCKIn Competition 2015, Makersfair'LISBON 2015.

Bibliography

City National Security website. http://citynationalsecurity.com/news/consectetur-adipiscing-elit/.

Context wiki definition. https://en.wiktionary.org/wiki/context.

2011. Application of surveillance tools to border surveillance - concept of operations. *online.*

Aarai, K., & Andrie, R. 2013. 3D Skeleton model derived from Kinect Depth Sensor Camera and its application to walking style quality evaluations. *In: International Journal of Advanced Research in Artificial Intelligence 2.*

Agarwal, A., & Triggs, B. 2004. 3D Human Pose from Silhouettes by Relevance Vector Regression. Computer Vision and Pattern Recognition.

Aggarwal, J., & Ryoo, M. 2011. Human activity analysis: A review. *ACM Computing Surveys*, **43**.

Aggarwal, J.K., & Park, S. 2004. Human motion: modeling and recognition of actions and interactions. 2nd International Symposium on 3D Data Processing, Visualization and Transmission 3DPVT.

Aitken, C., & Taroni, F. 19995. Statistics and the Evaluation of Evidence for Forensic Scientist.

Andersson, V. O., & Araujo, R. M. 2015. Person Identification Using Anthropometric and Gait Data from Kinect Sensor. *In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.*

Ariyanto, G., & Nixon, M.S. 2011. Model-based 3d gait biometrics. *In: In International Joint Conference on Biometrics (IJCB).*

Aziz, K.E., Merad, D., & Fertil, B. 2011a. People re-identification across multiple non-overlapping cameras system by appearance classification and silhouette part segmentation. 8th IEEE International Conference AVSS.

Aziz, K.E., Merad, D., & Fertil, B. 2011b. Person re-identification using appearance classification. *Pages 170–179 of: 8th International Conference on Image Analysis and Recognition, ICIAR.*

Bak, S., Corvee, E., Bremond, F., & Thonnat, M. 2010. Person re-identification using spatial covariance regions of human body parts. *Pages 435–440 of: Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on.*

Bak, S., Zaidenberg, S., Boulay, B., & Bremond, F. 2014. Improving person re-identification by viewpoint cues. *Pages 175–180 of: Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on.* IEEE.

Bak, S., Martins, F., & Bremond, F. 2015. Person re-identification by pose priors. *Pages 93990H–93990H of: SPIE/IS&T Electronic Imaging.* International Society for Optics and Photonics.

Baltieri, D., Vezzani, R., & Cucchiara, R. 2011a. 3dpes: 3d people dataset for surveillance and forensics. *Pages 59–64 of: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding.*

Baltieri, D., Vezzani, R., & Cucchiara, R. 2011b. Sarc3d: a new 3d body model for people tracking and re-identification. *Pages 197–206 of: Internat. Conf. on Image Analysis and Processing.* Springer.

Baltieri, D., Vezzani, R., & Cucchiara, R. 2015. Mapping appearance descriptors on 3d body models for people re-identification. *Internat. J. of Computer Vision*, **111**(3), 345–364.

Barbosa, B.I., Cristani, M., Bue, A. Del, Bazzani, L., & Murino, V. 2012a (October). Re-identification with RGB-D sensors. *In: First International Workshop on Re-Identification.*

Barbosa, I. B., Cristani, M., Alessio, D. B., Bazzani, L., & Murino, V. 2012b. Re-identification with RGB-D Sensors. *In: Computer Vision–ECCV 2012. Workshops and Demonstrations.*

Barrett, David. 2013 (July). *One surveillance camera for every 11 people in Britain, says CCTV survey.*

Bashir, K., Xiang, T., & Gong, S. 2009. Gait Representation Using Flow Fields. *In: British Machine Vision Conference (BMVC).*

Bashir, K., Xiang, T., & Gong, S. 2010. Gait recognition without subject cooperation. *Pattern Recognition Letters*, **31**, 2052–2060.

Bedagkar-Gala, A., & Shah, S.K. 2011. Multiple person re-identification using part based spatio-temporal color appearance model. *Pages 1721–1728 of: Computer Vision Workshops (ICCV Workshops).* IEEE.

Bedagkar-Gala, A., & Shah, S.K. 2014. Gait-Assisted Person Re-identification in Wide Area Surveillance. *Computer Vision - ACCV Workshops*, 633–649.

Belongie, S., Malik, J., & Puzicha, J. 2002. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence.

BenAbdelkader, C., Cutler, R.G., & Davis, L.S. 2004. Gait recognition using image self-similarity. *EURASIP Journal on Applied Signal Processing*, 572–585.

Benenson, R., Mathias, M., Timofte, R., & Gool, L. Van. 2012. Pedestrian detection at 100 frames per second. *CVPR.*

Benenson, R., Omran, M., Hosang, J., & Schiele, B. 2014. Ten Years of Pedestrian Detection, What Have We Learned? *Pages 613–627 of: Computer Vision - ECCV 2014 Workshops.*

Berclaz, J., Fleuret, F., Türetken, E., & Fua, P. 2011. Multiple Object Tracking using K-Shortest Paths Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 1806 – 1819.

Bertillon, A. 1896. Signaletic instructions including the theory and practice of anthropometrical identification. *The Werner Company.*

Bialkowski, A., Denman, S., Sridharan, S., Fookes, C., & Lucey, P. 2012. A database for person re-identification in multi-camera surveillance networks. *Pages 1– 8 of: Proceedings of the 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA 12).*

Bialkowski, A., Lucey, P., Wei, X., & Sridharan, S. 2013. Person Re-Identification Using Group Information. *Pages 1–6 of: International Conference on Digital Image Computing: Techniques and Applications (DICTA).*

Blake, R., & Shiffrar, M. 2007. Perception of human motion. *Annu. Rev. Psychol.*, **58**, 47–73.

Bobick, A.F., & Johnson, A.Y. 2001. Gait Recognition Using Static, Activity Specific Parameters. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR.

Bouchrika, I., Goffredo, M., Carter, J., & Nixon, M. 2011. On using gait in forensic biometrics. *Journal of Forensic Sciences*, **56**, 882–889.

Bouchrika, I., Carter, J.N., & Nixon, M.S. 2016. Towards automated visual surveillance using gait for identity recognition and tracking across multiple non-intersecting cameras. *Multimedia Tools and Applications*, **75**(2), 1201–1221.

Boyd, J.E., & Little, J.J. 2005. Biometric gait recognition. 19–42.

Cancela, B., Hospedales, T., & Gong, S. 2014. Open-World Person Re-Identification by Multi-Label Assignment Inference. *In: British Machine Vision Conference.*

Castro, F.M., Marín-Jimenez, M.J, & Medina-Carnicer, R. 2014. Pyramidal Fisher Motion for multiview gait recognition. *arXiv preprint arXiv:1403.6950.*

Chapelle, V., & Vapnik, V. 1999. Model Selection for Support Vector Machines. In Advances in Neural Information Processing Systems, Vol 12.

Chattopadhyay, P., Sural, S., & Mukherjee, J. 2015. Information fusion from multiple cameras for gait-based re-identification and recognition. *IET Image Processing*, **9**(11), 969–976.

Chen, C., Liang, J., Zhao, H., Hu, H., & Tian, J. 2009. Frame difference energy image for gait recognition with incomplete silhouettes. **30**, 997–984.

Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., & Murino, V. 2011. Custom Pictorial Structures for Re-identification. *Page 6 of: BMVC*, vol. 1.

Chunxiao, L., Shaogang, G., Loy, C., & Lin, X. 2012. Person Re-identification: What Features Are Important? *Pages 391–401 of: Computer Vision – ECCV 2012.*

Cortes, C., & Vapnik, V. 1995. Support-vector networks. *ML*.

Dalal, N., & Triggs, B. 2005. Histograms of Oriented Gradients for Human Detection. *In: Computer Vision and Pattern Recognition*.

Dalal, N., Triggs, B., & Schmid, C. 2006. Human Detection Using Oriented Histograms of Flow and Appearance. *In: 9th European Conference on Computer Vision*.

Dantcheva, A., Velardo, C., D'angelo, A., Dugelay, & Jean-Luc. 2010. Bag of soft biometrics for person identification: New trends and challenges. Mutimedia Tools and Applications, Springer.

Ding, Y., Meng, X., Chai, G., & Tang, Y. 2011. User identification for instant messages. *Pages 113–120 of: International Conference on Neural Information Processing*. Springer.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. 2009. Pedestrian detection: A benchmark.

Dollár, Piotr. *Caltech Pedestrian Detection evaluation code*. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/DollarEvaluationCode.

Dollár, Piotr. *Piotr's Image and Video Matlab Toolbox (PMT)*. http://vision.ucsd.edu~/pdollar/toolbox/doc/index.html.

Doretto, G., Sebastian, T., Tu, P., & Rittscher, J. 2011. Appearance-based person reidentification in camera networks: Problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, **2**, 127–151.

Eastlack, M.E., Arvidson, J., Snyder-Mackler, L., Danoff, J.V., & McGarvey, C.L. 1991. Interrater reliability of videotaped observational gait-analysis assessments. *Physical Therapy*, **71**, 465–472.

Ess., A., Leibe., B., & Van Gool., L. 2007. Depth and Appearance for Mobile Scene Analysis. *ICCV*.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. 2010. Person Re-Identification by Symmetry-Driven Accumulation of Local Featuresn. *Pages 2360 – 2367 of: IEEE Conference onComputer Vision and Pattern Recognition (CVPR)*.

Fernández, D.L., Madrid-Cuevas, F.J., Carmona-Poyato1, A., Muñoz-Salinas, R., & Medina-Carnicer, R. 2016. A new approach for multi-view gait recognition on unconstrained paths. *Journal of Visual Communication and Image Representation*, **38**, 396–406.

Ferryman, J., & Shahrokni, A. 2009. An overview of the PETS 2009 challengen. *In: 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*.

Figueira, D., Bazzani, L., Minh, H.Q., Cristani, M., Bernardino, A., & Murino, V. 2013. Semisupervised multi-feature learning for person reidentification. *Pages 111 – 116 of: 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*.

Figueira, D., Taiana, M., Nambiar., A., Nascimento, J.C., & Bernardino, A. 2014. The HDA+ Data Set for Research on Fully Automated Re-identification Systems. *Pages 241–255 of: roc. of ECCV 2014 Workshop on Visual Surveillance and Re-Identification*.

Forsyth, D.A., & Ponce, J. 2002. Computer Vision: A Modern Approach. *Prentice Hall.*

Gabel, M., Gilad-Bachrach, R., Renshaw, E., & Schuste, A. 2012. Full Body Gait Analysis with Kinect. *In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).*

Gala, A.B., & Shah, S.K. 2014. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, **32**, 270–286.

Garcia, J., Martinel, N., Micheloni, C., & Gardel, A. 2015. Person re-identification ranking optimisation by discriminant context information analysis. *Pages 1305–1313 of: Proceedings of the IEEE International Conference on Computer Vision.*

Gavrila, D. 1999a. The visual analysis of human movement: A survey. *Comput. Vis. Image Understand.*, **73**, 82–98.

Gavrila, D.M. 1999b. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, **73**, 82–98.

Geiger, J.T., Kneißl, M., Schuller, B.W., & Rigoll, G. 2014. Acoustic gait-based person identification using hidden Markov models. *Pages 25–30 of: Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop.* ACM.

Gianaria, E., Grangetto, M., Lucenteforte, M., & Balossino, N. 2014. Human Classification Using Gait Features. *Biometric Authentication*, **8897**, 16–27.

Goffredo, M., Carter, J., & Nixon, M. 2008. Front-view Gait Recognition. *In: Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems.*

Gong, S., Cristani, M., Loy, C.C., & Hospedales, T.M. 2014. The Re-identification Challenge. *erson Re-Identification*, 1–20.

Gordon, C., Churchill, T., & Clauser, C.E. 1989. 1988 anthropometric survey of u.s. army personnel: Methods and summary statistics. *United States Army Natick Research Tecnical report.*

Gowsikhaa, D., Abirami, S., & Baskaran, R. 2014. Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, **42**, 747–765.

Gray, D., & Tao, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *Pages 262–275 of: European Conf. on computer vision.* Springer.

Gray, D., Brennan, S., & Tao, H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. *In: Proc. IEEE Internat. Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3.

Gross, R., & Shi, J. 2001. The cmu motion of body (mobo) database. *Technical Report CMU-RI-TR-01- 18, Robotics Institute, Carnegie Mellon University.*

Grother, P., & Phillips, P.J. 2004. Models of large population recognition performance. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Hamdoun, O., Moutarde, F., Stanciulescu, B., & Steux, B. 2008. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. *In: 2nd ACM/IEEE International Conference on Distributed Smart Cameras.*

Hampapur, A., Brown, L., Connell, J., Pankanti, S., Senior, A., & Tian, Y. 2003. Smart surveillance: applications, technologies and implications. *Page 1133–1138 of: IEEE Pacific-Rim Conference On Multimedia.*

Han, J., & Bhanu, B. 2006. Individual recognition using gait energy image. *Pattern Analysis and Machine Intelligence*, **28**, 316–322.

Harold, W.K. 1955. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, **2**, 83–97.

Hirzer, M., Beleznai, C., Roth, P.M, & Bischof, H. 2011. Person re-identification by descriptive and discriminative classification. *Pages 91–102 of: Scandinavian Conf. on Image analysis.* Springer.

Hofmann, M., & Rigoll, G. 2012. Improved Gait recognition using gradient histogram energy image. *In: IEEE International Conference on Image Processing (ICIP).*

Hofmann, M., Sural, S., & Rigoll, G. 2011. Gait recognition in the presence of occlusion: A new dataset and baseline algorithms. *Pages 99–104 of: 19th Internat. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision.*

Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., & Rigoll, G. 2014. The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recognition of Subjects and Traits. *J. of Visual Communication and Image Representation, Special Issue on Visual Understanding and Applications with RGB-D Cameras*, **25**, 195–206.

Hosang, J., Omran, M., Benenson, R., & Schiele, B. 2015. Taking a Deeper Look at Pedestrians. *Pages 4073–4082 of: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Hu, D., Pan, S., Zheng, V., Liu, N., & Yang, Q. 2008. Real world activity recognition with multiple goals. *Page 30–39 of: in Proc. ACM 10th Int. Conf. Series.*

Hu, W., Tan, T., Wang, L., & Maybank, S. 2004a. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics*, **34**, 334 – 352.

Hu, W., Tan, T., Wang, L., & Maybank, S. 2004b. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst. Cybern. C, Appl. Rev.*, **34**, 334–352.

Iwashita, Y., Baba, R., Ogawara, K., & Kurazume, R. 2010. Person Identification from Spatio-temporal 3D Gait. *In: Proceedings of the International Conference on Emerging Security Technologies.*

Iwashita, Y., Ogawarab, K., & Kurazume, R. 2014. Identification of people walking along curved trajectories. *Pattern Recognition Letters*, **48**, 60–69.

Johansson, G. 1973. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, **14**, 201–211.

John, V., Englebienne, G., & Krose, B. 2013. Person re-identification using height-based gait in colour depth camera. *Pages 3345–3349 of: 2013 IEEE Internat. Conf. on Image Processing.* IEEE.

Johnson, A.Y., & Bobick, A.F. 2001. A multi-view method for gait recognition using static body parameters. *Pages 301–311 of: Internat. Conf. on Audio-and Video-Based Biometric Person Authentication.* Springer.

Josinski, H., Michalczuk, A., Kostrzewa, D., Świtoński, A., & Wojciechowski, K. 2014. Heuristic Method of Feature Selection for Person Re-identification Based on Gait Motion Capture Data. 6th Asian Conference- ACIIDS.

Jungling, K., & Arens, M. 2010. Local Feature Based Person Reidentification in Infrared Image Sequences. *Page 448–455 of: In Proc. of IEEE Conf. on Advanced Video and Signal-Based Surveillance.*

Kale, A., Cuntoor, N., Yegnanarayana, B., Rajagopalan, A.N., & Chellappa, R. 2003. Gait analysis for human identification. *Pages 706–714 of: Proceedings of the 4th international conference on Audio- and video-based biometric person authentication AVBPA'03.*

Kawai, R., Makihara, Y., Hua, C., Iwama, H., & Yagi, Y. 2012. Person re-identification using view-dependent score-level fusion of gait and color features. *Pages 2694–2697 of: Pattern Recognition (ICPR), 2012 21st Internat. Conf. on.* IEEE.

Kress, T., & Daum, I. 2003. Developmental prosopagnosia: A review. *Behavioural neurology*, **14**(3-4), 109–121.

Kressig, R.W., & Beauchet, O. 2006. Guidelines for clinical applications of spatiotemporal gait analysis in older adults. *Aging Clinical and Experimental Research*, **18**, 174–176.

Kusakunniran, W., Wu, Q., Li, H., & Zhang, J. 2009. Multiple views gait recognition using view transformation model based on optimized gait energy image. *Pages 1058–1064 of: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th Internat. Conf. on.* IEEE.

Kviatkovsky, I., Adam, A., & Rivlin, E. 2013. Color invariants for person reidentification. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Lam, T.H.W., Cheung, K.H., & Liu, J.N.K. 2011. Gait flow image: A silhouette-based gait representation for human identification. *In: Pattern Recognition.*

Layne, R., Hospedales, T., & Gong, S. 2012. Person Re-identification by Attributes. *British Machine Vision Conference.*

Lee, L., & Grimson, W. 2002. Gait Analysis for Recognition and Classification. *Pages 155–162 of: Proc. IEEE International Conference on Automatic Face and Gesture Recognition.*

Lee, T.K.M., Belkhatir, M., & Sanei, S. 2014. A comprehensive review of past and present vision-based techniques for gait recognition. *Multimedis Tools Applications*, **72**, 2833–2869.

Leibe, B., Seemann, E., & Schiele, B. 2005. Pedestrian Detection in Crowded Scenes. Computer Vision and Pattern Recognition.

Leng, Q., Hu, R., Liang, C., Wang, Y., & Chen, J. 2015. Person re-identification with content and context re-ranking. *Multimedia Tools and Applications*, **74**(17), 6989–7014.

Liao, S., Mo, Z., Zhu, J., Hu, Y., & Li, S.Z. 2014. Open-set Person Re-identification. *arXiv preprint arXiv:1408.087*.

Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., & Zingaretti, P. 2016. Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration. *Pages 1–11 of: International Workshop on Face and Facial Expression Recognition from Real World Videos*. Springer.

Liu, C., Gong, S., Loy, C.C., & Lin, X. 2012. Person Re-identification: What Features Are Important? *Pages 391–401 of: European Conference on Computer Vision, International Workshop on Re-Identification*.

Liu, H., Hu, L., & Ma, L. 2017. Online RGB-D person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology*, **2**(1), 48–55.

Loudon, J. 2008. The clinical orthopedic assessment guide. *Kansas: Human Kinetics*.

Lowe, D.G. 2004. Distinctive Image Features from Scale Invariant Features. International Journal of Computer Vision.

Loy, C.C., Xiang, T., & Gong, S. 2009. Multi-camera activity correlation analysis. *Pages 1988–1995 of: Computer Vision and Pattern Recognition*. IEEE.

Lucas, B.D., & Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. *In: Proceedings of Imaging Understanding Workshop*.

Luo, P., Tian, Y., Wang, X., & Tang, X. 2014. Switchable Deep Network for Pedestrian Detection. *Columbus, OH*, 899 – 906.

López-Fernández, D., Madrid-Cuevas, F.J., Carmona-Poyato, A., Marín-Jiméne, M.J., & Muñoz-Salinas, R. 2014. The AVA Multi-View Dataset for Gait Recognition. *Pages 26–39 of: Activity Monitoring by Multiple Distributed Sensing, Lecture Notes in Computer Science*.

Makihara, Y., Matovski, D.S, Nixon, M.S., Carter, J.N., & Yagi, Y. 2015. Gait recognition: Databases, representations, and applications. *Wiley Encyclopedia of Electrical and Electronics Engineering*.

Maloney, L.T., & Wandell, B.A. 1986. Color constancy: a method for recovering surface spectral reflectance. *Journal of the Optical Society of America*, 29–33.

Middleton, L., Buss, A.A., Bazin, A.I., & Nixon, M.S. 2005. A floor sensor system for gait recognition. *Pages 171 – 176 of: Fourth IEEE Workshop on Automatic Identification Advanced Technologies*.

Moeslund., T.B., Hilton, A., & Krüger, V. 2006. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, **104**, 90–126.

Moon, H., & Phillips, P.J. 2001. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception*, **30**, 303–321.

Moreno, P., Figueira, D., Bernardino, A., & Victor, J.S. 2015. People and Mobile Robot Classification Through Spatio-Temporal Analysis of Optical Flow. *International Journal of Pattern Recognition and Artificial Intelligence*.

Munaro, M., Ghidoni, S., Dizmen, D.T., & Menegatti, E. 2014a. A feature-based approach to people re-identification using skeleton keypoints. *Pages 5644–5651 of: Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE.

Munaro, M., Fossati, A., Basso, A., Menegatti, E., & Van Gool, L. 2014b. One-shot person re-identification with a consumer depth camera. *Pages 161–181 of: Person Re-Identification*. Springer.

Muramatsu, D., Makihara, Y., & Yagi, Y. 2014. View transformation-based cross-view gait recognition using transformation consistency measure. *In: International Workshop on Biometrics and Forensics (IWBF)*.

Nambiar, A., Taiana, M., Figueira, D., Nascimento, J.C., & Bernardino, A. 2014a. A Multi-camera video data set for research on High-Definition surveillance. International Journal of Machine Intelligence and Sensory Signal Processinge.

Nambiar, A., Taiana, M., Figueira, D., Nascimento, J.C., & Bernardino, A. 2014b. A multi-camera video dataset for research on high-definition surveillance. *International Journal of Machine Intelligence and Sensory Signal Processing*.

Nambiar, A., Bernardino, A., & Nascimento, J.C. 2015. Shape Context for soft biometrics in person re-identification and database retrieval. *Pattern Recognition Letters*, **68**, 297–305.

Nambiar, A., Nascimento, J.C., & Bernardino, A. 2016a. Gait based Person Re-identification: a Survey. *Image and Vision Computing (submitted)*.

Nambiar, A., Bernardino, A., & Nascimento, J.C. 2016b. Person Re-identification based on Human query on Soft Biometrics using SVM regression. *In: 1th International Conference on Computer Vision Theory and Applications*.

Nambiar, A, Nascimento, J.C., Bernardino, A., & Santos-Victor, J. 2016c. Person Re-identification in Frontal Gait Sequences via Histogram of Optic Flow Energy Image. *Pages 250–262 of: Internat. Conf. on Advanced Concepts for Intelligent Vision Systems*. Springer.

Nambiar, A., Bernardino, A., Nascimento, J.C., & Fred, A. 2017a. Context-Aware Person Re-identification in the Wild via fusion of Gait and Anthropometric features. *In: B-WILD Workshop at 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*.

Nambiar, A., Bernardino, A., Nascimento, J.C., & Fred, A. 2017b. Towards view-point invariant Person Re-identification via fusion of Anthropometric and Gait Features from Kinect measurements. *In: International Conference on Computer Vision Theory and Applications (VISAPP), Porto*.

Nambiar, A., Bernardino, A., Nascimento, J.C., Fred, A., & Santos-Victor, J. 2017 **(In preparation)**. Context for Person Re-identification: An ensemble fusion Re-ID framework leveraging soft biometric features and its extension towards cross-context analysis. *In: Special Issue on Biometrics in the Wild, Image and Vision Computing.*

Nambiar, A.M., Correia, P.L., & Soares, L.D. 2012. Frontal Gait Recognition Combining 2D and 3D Data. *In: Proceedings of the on Multimedia and Security.* MM&Sec '12.

Nascimento, J.C., Figueiredo, M.A.T., & Marques, J.S. 2013. Activity Recognition Using A Mixture of Vector Fields. **22**.

Niu, W., Jiao, L., Han, D., & Wang, Y. F. 2003. Real-time multiperson tracking in video surveillance. *Pages 1144 – 1148 of: Fourth Pacific Rim Conference on Multimedia, Information, Communications and Signal Processing.*

Nixon, M.S., & Carter, J.N. 2006. Automatic Recognition by Gait. *Proceedings of the IEEE*, **94**, 2013 – 2024.

Nixon, M.S, Tan, T., & Chellappa, R. 2010. *Human identification based on gait.* Vol. 4. Springer Science & Business Media.

Nixon, M.S., Correia, P.L., Nasrollahi, K., Moeslund, T.B., Hadidd, A., & Tistarelli, M. 2015. On soft biometrics. **68**, 218–230.

Padole, C., & Proença, H. 2017. An aperiodic feature representation for gait recognition in cross-view scenarios for unconstrained biometrics. *Pattern Analysis and Applications*, 73–86.

Pala, F., Satta, R., Fumera G., & Roli, F. 2015. Multi-modal Person Re-Identification Using RGB-D Cameras. *IEEE Trans. on Circuits and Systems for Video Technology*, **26**(4), 788 – 799.

Palmisano, C., Tuzhilin, A., & Gorgoglione, M. 2008. Using context to improve predictive modeling of customers in personalization applications. *IEEE transactions on knowledge and data engineering*, **20**(11), 1535–1549.

Panniello, U, Hill, S, & Gorgoglione, M. 2016. Using context for online customer re-identification. *Expert Systems with Applications*, **64**, 500–511.

Perlin, M.W. 2010. Explaining the Likelihood Ratio in DNA Mixture Interpretation. *In: In the Proceedings of Promega's Twenty First International Symposium on Human Identification.*

Phillips, P., Sarkar, S., Robledo, I., Grother, P., & Bowyer, K. 2002. Baseline results for the challenge problem of human id using gait analysis. *Page 130–135 of: Fifth IEEE International Conference on Automatic Face and Gesture Recognition.*

Phillips, P.J., Moon, H., Rizvi, S.A., & Rauss, P.J. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 1090–1104.

Plantinga, Alvin. 1961. Things and Persons. *Review of Metaphysics*, **14**(March), 493–519.

Pohjalainen, J., Räsänen, O., & Kadioglu, S. 2015. Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Computer Speech & Language*, **29**(1), 145–171.

Reid, D., & Nixon, M. 2011. Using comparative human descriptions for soft biometrics. *In: In The first International Joint Conference on Biometrics.*

Reid, D.A., Nixon, M.S., & Stevenage, S.V. 2014. Soft Biometrics; Human Identification Using Comparative Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**, 1216–1228.

Rhodes, H.T.F. 1956. Alphonse Bertillon: Father of Scientific Detection.

Riccio, D., Marsico, M.D., Distasi, R., & Ricciardi, S. 2014. A comparison of approaches for person re-identification. *Pages 189–198 of: International Conference on Pattern Recognition Applications and Methods, ESEO.*

Robertson, N., & Reid, I. 2006a. A general method for human activity recognition in video. **104**, 232–248.

Robertson, N., & Reid, I. 2006b. A general method for human activity recognition in video. *Computer Vision and Image Understanding*, **104**, 232 – 248.

Ross, A. 2007. An introduction to multibiometrics. *In: 15th European Signal Processing Conference.*

Ross, A., & Jain, A.K. 2007. Human recognition using biometrics: an overview. *Annales Des Télécommunications*, 11–35.

Ross, A.A., Nandakumar, K., & Jain, A. 2006. *Handbook of multibiometrics.* Vol. 6. Springer Science & Business Media.

Samangooei, S., & Nixon, M.S. 2010. Performing Content-based Retrieval of Humans using Gait Biometrics. *Multimedia Tools and Applications*, **49**, 195–212.

Samangooei, S., Nixon, M., & b. Guo. 2008. The use of semantic human description as a soft biometric. In Proceedings of BTAS.

Sarkar, S., Phillips, P.J., Liu, Z., Vega, I. R., Grother, P., & Bowyer, K.W. 2005. The human id gait challenge problem: data sets, performance, and analysiss. *EEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 162–177.

Satta, R., Pala, F., Fumera, G., & Roli, F. 2014. People Search with Textual Queries About Clothing Appearance Attributes. *Person Re-Identification*, 371–389.

Seely, R.D., Samangooei, S., Middleton, L., Carter, J. N., & Nixon, M. S. 2008. The university of southampton multi-biometric tunnel and introducing a novel 3d gait dataset. *In: 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems BTAS.*

Shi, Z., Hospedales, T.M., & Xiang, T. 2015. ransferring a Semantic Representation for Person Re-Identification and Search. *Pages 4184–4193 of: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Shitrit, H.B., Berclaz, J., Fleuret, F., & Fua, P. 2014. Multi-Commodity Network Flow for Tracking Multiple People. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**, 1614–1627.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., & Moore, R. 2013. Real-time human pose recognition in parts from single depth images. *In: Communications of the ACM (CACM), 56(1).*

Siebel, N.T., & Maybank, S.J. 2002. Fusion of Multiple Tracking Algorithms for Robust People Tracking. *Pages 373–387 of: Proceedings of the 7th European Conference on Computer Vision ECCV.*

Silva, H., & Fred, A. 2007. Feature subspace ensembles: a parallel classifier combination scheme using feature selection. *Pages 261–270 of: International Workshop on Multiple Classifier Systems.* Springer.

Sivapalan, S. 2014. Human Identification from Video Using Advanced Gait Recognition Techniques. *PhD thesis, Queensland University of Technology.*

Sivapalan, S., Chen, D., Denman, S., Sridharan, S., & Fookes, C. 2011. 3D Ellipsoid Fitting for Multiview Gait Recognition. *Pages 355–360 of: In Proceedings of 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS).*

Sivapalan, S., Chen, D., Denman, S., Sridharan, S., & Fookes, C. 2012. The Backfilled GEI - A Cross-capture Modality Gait Feature for Frontal and Side-view Gait Recognition. *In: International Conference on Digital Image Computing Techniques and Applications (DICTA).*

Smola, A., & Schölkopf, B. 1998. A Tutorial on Support Vector Regression. NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK.

Smola, A., & Schölkopf, B. 2004. A tutorial on support vector regression. *In: Statistics and Computing 14 (3): 199–222.*

Sridharan, K., Nayak, S., Chikkerur, S., & Govindaraju, V. 2005. A probabilistic approach to semantic face retrieval system. *in Audio-and video-based biometric person authentication, Springer.*

Stevenage, S.V., Nixon, M.S., & Vince, K. 1999. Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology*, **13**, 513– 526.

Sumi, Shigemasa. 1984. Upside-down presentation of the Johansson moving light-spot pattern. *Perception*, **13**(3), 283–286.

Taiana, M., Figueira, D., Nambiar, A., Nascimento, J., & Bernardino, A. 2014. Towards Fully Automated Person Re-Identification. *Pages 140 – 147 of: 9th International Conference on Computer vision Theory and Applications VISAPP.*

Thompson, P.D., & Nutt, J.G. 2012. Gait disorders. *Bradley's Neurology in Clinical Practice*, **22**.

Truong, D.N.T., Achard, C., & Khoudour, L. 2010. People re-identification by classification of silhouettes based on sparse representation. International Conference on Image Processing Theory Tools and Applications (IPTA).

Tu, P., Doretto, G., Krahnstoever, N., Perera, A.G.A., Wheeler, F., Liu, X., Rittscher, J., Sebastian, T., Yu, T., & K.Harding. 2007. An intelligent video framework for homeland protection. Proceedings of SPIE Defence and Security Symposium—Unattended Ground, Sea, and Air Sensor Technologies and Applications.

Vapnik. 1995. The Nature of Statistical Learning Theory. *In: Springer.*

Vezzani, R., Baltieri, D., & Cucchiara, R. 2013. People Re-identification in Surveillance and Forensics: a Survey. *CM Computing Surveys CSUR*, **46**, 1–36.

Wang, C., Zhang, J., Wang, L., Pu, J., & Yuan, X. 2012. Human Identification Using Temporal Information Preserving Gait Template. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **34**, 2164–2176.

Wang, J., She, M., Nahavandi, S., & Kouzani, A. 2010. A review of vision-based gait recognition methods for human identification. International Conference on Digital Image Computing: Techniques and Applications (DICTA).

Wang, L., Tieniu, T., H.Weiming, & Huazhong, N. 2003a. Automatic gait recognition based on statistical shape analysis. *EEE Transactions on Image Processing*, **12**, 1120 – 1131.

Wang, L., Tan, T., Ning, H., & Hu, W. 2003b. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1505 – 1518.

Wang, T., Gong, S., Zhu, X., & Wang, S. 2014. Person re-identification by video ranking. *Pages 688–703 of: European Conf. on Computer Vision*. Springer.

Wang, T., Gong, S., Zhu, X., & Wang, S. 2016. Person Re-Identification by Discriminative Selection in Video Ranking. *IEEE Trans. on PAMI*, **38**(12).

Wang, X., Doretto, G., Sebastian, T., Rittscher, J., & PeterTu. 2007. Shape and appearance context modeling. IEEE 11th ICCV.

Wei, L., Tian, Y., Wang, Y., & Huang, T. 2015. Swiss-System Based Cascade Ranking for Gait-Based Person Re-Identification. *Pages 197–202 of: Twenty-Ninth AAAI Conf. on Artificial Intelligence*.

Whitney, A Wayne. 1971. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, **100**(9), 1100–1103.

Whittle, Michael W. 1996. Clinical gait analysis: A review. *Human Movement Science*, **15**(3), 369–387.

Wiegler, L. 2008. Big brother in the big apple [national security - video surveillance]. *Engineering & Technology*, **3**, 24–27.

Wu, Z., Li, Y., & Radke, R.J. 2015. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(5), 1095–1108.

Yang, Y., Tu, D., & Li, G. 2014. Gait Recognition Using Flow Histogram Energy Image. *In: International Conference on Pattern Recognition.*

Zhang, C., & Viola, P. 2007. Multiple-Instance Pruning For Learning Efficient Cascade Detectors. *NIPS.*

Zhang, E., Zhao, Y., & Xiong, W. 2010. Active energy image plus 2DLPP for gait recognition. *Signal Processing*, **90**, 2295–2302.

Zhang, L., Kalashnikov, D.V., Mehrotra, S., & Vaisenberg, R. 2014. Context-based person identification framework for smart video surveillance. *Machine Vision and Applications*, **25**(7), 1711–1725.

Zhang, Y., Wu, X., & Ruan, Q. June 2009. Combining procrustes shape analysis and shape context descriptor for silhouette-based gait recognition. Electronics Letters.

Zhao, G., Liu, G., Li, H., & Pietikäinen, M. 2006. 3D gait recognition using multiple cameras. *In: 7th International Conference on Automatic Face and Gesture Recognition.*

Zhao, R., Ouyang, W., & Wang, X. 2013. Unsupervised salience learning for person re-identification. *Pages 3586–3593 of: IEEE Conf. on Computer Vision and Pattern Recognition.*

Zheng, L., Yang, Y., & Hauptmann, A.G. 2016. Person Re-identification: Past, Present and Future. *CoRR*, **arXiv:1610.02984**.

Zheng, W.S., Gong, S., & Xiang, T. 2009. Associating Groups of People. *Pages 23.1–23.11 of: Proc. BMVC.*

Zweig, M.H., & Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**, 561–577.