Cross-Modal and Multi-Modal Person Re-identification with RGB-D Sensors

（RGB-Dセンサを用いた複数モダリティの相互利用に基づく人物同定）

Md Kamal Uddin

Graduate School of Science and Engineering

Saitama University, Japan

Supervisor: Professor Yoshinori Kobayashi

A thesis submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

September, 2021

To my parents
with my deepest gratitude

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, *Professor Yoshinori Kobayash* and my ex-supervisor, *Professor Yoshinori Kuno*, Saitama University, Japan, for their immense support and guidance during my Ph.D studies. Under their supervision, I learnt how to define and solve challenging problems as an individual researcher.

I am incredibly grateful to *Dr. Antony Lam* for his prompt review comments on the articles, valuable advice and feedback during my Ph.D studies. I am forever thankful to *Assoc. Professor Hisato Fukuda* for his continuous support in the laboratory providing all necessary devices to perform experiments.

I am also grateful to my thesis examiners, *Professor Yoshinori Kobayashi*, *Professor Tetsuya Shimamura*, *Professor Jun Ohkubo* and *Professor Takashi Komuro*, for their valuable comments and advice to complete my doctoral studies.

I would like to thank all of my lab members, specially *Kouyou Otsu*, *Mahmudul Hasan* and *Matiqul Islam*, for their cooperation and unconditional support. I would like to extend my gratitude to all students who participated in my experimental video data.

The financial support from the Government of Japan's Ministry of Education, Culture, Sports, Science and Technology (Monbukagakusho:MEXT) has made it possible to pursue my Ph.D study. In this regard, I would like to thank Bangladesh Government, UGC (University Grants Commission of Bangladesh) and NSTU (Noakhali Science and Technology University) to grant my study leave for Ph.D study.

Finally, special thanks goes to my beloved wife, *Nishat Sharmin*, for her unconditional love, undying support and understanding, my daughter, *Safeeya Kamal Yasha*, for her cute actions that have helped me cheerful during my academic life.

Abstract

Person re-identification (Re-ID) is one of the most important tools of intelligent video-surveillance systems, which aims to recognize an individual across different non-overlapping sensors of a camera network. It is a very challenging task in computer vision because the visual appearance of an individual changes due to the variations in viewing angle, illumination intensity, pose, occlusion and diverse cluttered background. The general objective of this thesis is to tackle some of these constraints by proposing different approaches, which exploit modern RGB-D sensor-based additional information.

At first, we present a novel cross-modal person re-identification technique by exploiting local shape information of an individual, which bridges the domain gap between two modalities (RGB and Depth). The core idea is, most of the existing Re-ID systems widely use RGB-based appearance cues, which is not suitable when lighting conditions are very poor. However, for many security reasons, sometimes continued surveillance via camera in low lighting conditions is inevitable. To overcome this problem, we take advantage of the depth sensor based cameras (e.g. Microsoft Kinect and Intel RealSense Depth camera), which can be installed in dark places to capture video, while RGB based cameras can be installed in good lighting conditions. Such types of heterogeneous camera networks can be advantages due to the different sensing modalities available but face challenges to recognize people across depth and RGB cameras. In this approach, we propose a body partitioning method and novel HOG based feature extraction technique on both modalities, which extract local shape information from regions within an image. We find that combining the estimated features on both modalities can sometimes help to better reduce visual ambiguities of appearance features caused by lighting conditions and clothes. We also exploit an effective metric learning approach which obtains a better re-identification accuracy across RGB and depth domain.

In this dissertation, we also present two novel multi-modal person re-identification methods. In the first method, we introduce a depth guided attention-based person re-identification method in multi-modal scenario, which takes into account the depth-based additional information in the form of an attention mechanism. Most of the existing methods rely on complex dedicated attention-based architecture for feature fusion and thus become unsuitable for real-time deployment. In our approach, we propose a depth-guided foreground extraction mechanism that helps the model to dynamically select the more relevant convolutional filters of the backbone CNN architecture, for enhanced feature representation and inference.

In our second method, we propose a novel person re-identification technique that exploits the advantage of using multi-modal data for fusing in dissimilarity space, where we design a 4-channel RGB-D image input in the Re-ID framework. Additionally, lack of a proper RGB-D Re-ID dataset prompts us to collect a new RGB-D Re-ID dataset named SUCVL RGBD-ID, including RGB and depth images of 58 identities from three cameras where one camera was installed in poor illumination conditions and the remaining two cameras were installed in two different indoor locations with different indoor lighting environments.

Finally, extensive experimental evaluations on our dataset and publicly available datasets demonstrate that our proposed methods are efficient and outperform all the related state-of-the-art methods.

Keywords: *Video surveillance, Person re-identification, RGB-D sensors, Cross-modal person re-identification, Heterogeneous camera network, Multi-modal person re-identification, Depth guided attention, Dissimilarity space, Triplet loss.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In recent years, from the security and forensics concerns, the deployment of Closed-Circuit TeleVision cameras (CCTV) has increased exponentially in both public and private areas including supermarkets, shopping complexes, airports, railway stations, university campuses, housing apartments and workplaces. According to a study by Cisco, video surveillance traffic is projected to increase tenfold between 2015 and 2020 [1]. As the volume of surveillance video has increased exponentially, making the continuous monitoring of surveillance data is quite impossible for a human operator due to lack of attention. Moreover, there is foremost possibility of unexpected event can take place simultaneously in multiple cameras when a large camera network (hundreds or thousands of cameras) is installed in a large area (e.g. airport). Therefore, it is also impossible to monitor all unexpected simultaneous events because a human operator can concentrate only on one particular camera or event.

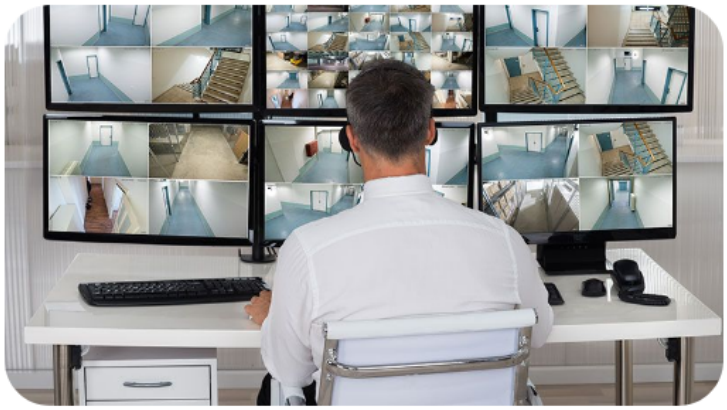

Figure 1.1: A security officer monitoring the CCTV [2].

Fig. 1.1 shows a typical scenario of a security officer monitoring a surveillance camera network, in which he can report alarms manually while observing any unexpected event in the scene. There is also have a privacy issue concerning the misuse of technologies, which is ongoing debate, while depending on human operator. Thus, it is essential to develop intelligent video surveillance systems that could provide informative data to the human operator and draw its attention whenever it is required. Intelligent surveillance system requires the ability to track or associate people across multiple cameras. The recognizing and associating the same person in a distributed multi-cameara system is known in the computer vision and pattern recognition community as person re-identification. Considering the various practical applications and challenges, it has been a unique and interesting field in research communities and industries in the last few years.

1.2 Person Re-identification

Person re-identification (Re-ID) is an important video-surveillance task of recognizing an individual over a set of disjoint camera views. Fig. 1.2 shows a schematic illustration of the re-identification problem where a set of probe images is captured from one camera view and a set of gallery images is captured from another disjoint camera views, a person re-identification system attempts to match each instance from the probe set against all from the gallery set.

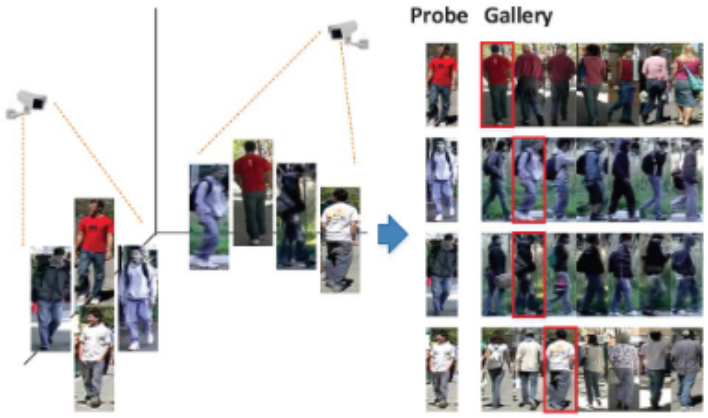

Figure 1.2: Schematic representation of person re-identification problem. In the left side, images are captured from disjoint camera views. In the right side, each row contains a probe image, and the corresponding rank gallery set, where the true match is marked in red box [3].

Person re-identification problems can be divided into two categorizes depending on the number of frames available of each person in the probe and gallery sets. If only one frame of each person is available in both the probe and gallery sets then it is called single-shot, and if multiple frames available in both sets then it is called multi-shot. Even though the multi-shot case provides more information, its computational cost is high over single-shot.

Person re-identification can be further subdivided into two other categories: long-term and short-term person re-identification. When the pedestrians keep their clothes unchanged while passing across the disjoint cameras in a short period, this scenario is called the short-term problem. However, this is not always true in practice because pedestrians are high likely to reappear after a long period , such as several days, the re-identification scenario is called the long-term problem.

Moreover, different sensory data such RGB, Depth and Skeleton information, acquired by modern RGB-D sensors, which can be used combinedly to construct robust features. When RGB data are combined with depth or skeleton information to improve the Re-ID performance, it refers as **multi-modal person re-identification**. In many real-world scenario, when matching RGB with depth modalities is important, for example, a video surveillance system that must recognize the individuals in poorly illuminated environments, which refers as **cross-modal person re-identification**.

Person re-identification has several practical applications to forensic search, multi-camera tracking, access control, sports analytics in a collection of video sequences. Recently it has also been applied to service robots and human-robot interaction for elderly monitoring and assistance to perform personalized tasks [8]. Another practical application for cross-modal Re-ID is autonomous self-driving vehicles, which require tracking pedestrians around their vicinity, where some regions are covered by LiDAR sensors, and others by RGB cameras [37].

1.3 Challenges of Person Re-ID

Though person re-identification has been studied over the past years, it is a still challenging task in computer vision community and remains unsolved. The major challenges in person re-identification are:

Pose and viewpoint variations: One of the biggest challenges is the variation in pose and appearance of the person in various cameras and in time as shown in Fig 1.3 (a). When the people are observed under the different camera views, the same

individual may look different (intra-class variation) and different individuals may look similar (inter-class variation). This variability happens due to the different camera orientation and changes in subject pose, camera resolution or visual appearance of the person itself. In Fig. 1.3 (b) shows the same person's images taken from different cameras. Even though the person is wearing same clothes across the camera, the color of their clothing looks significantly different in images. These intra-class variation makes the Re-ID more challenging.

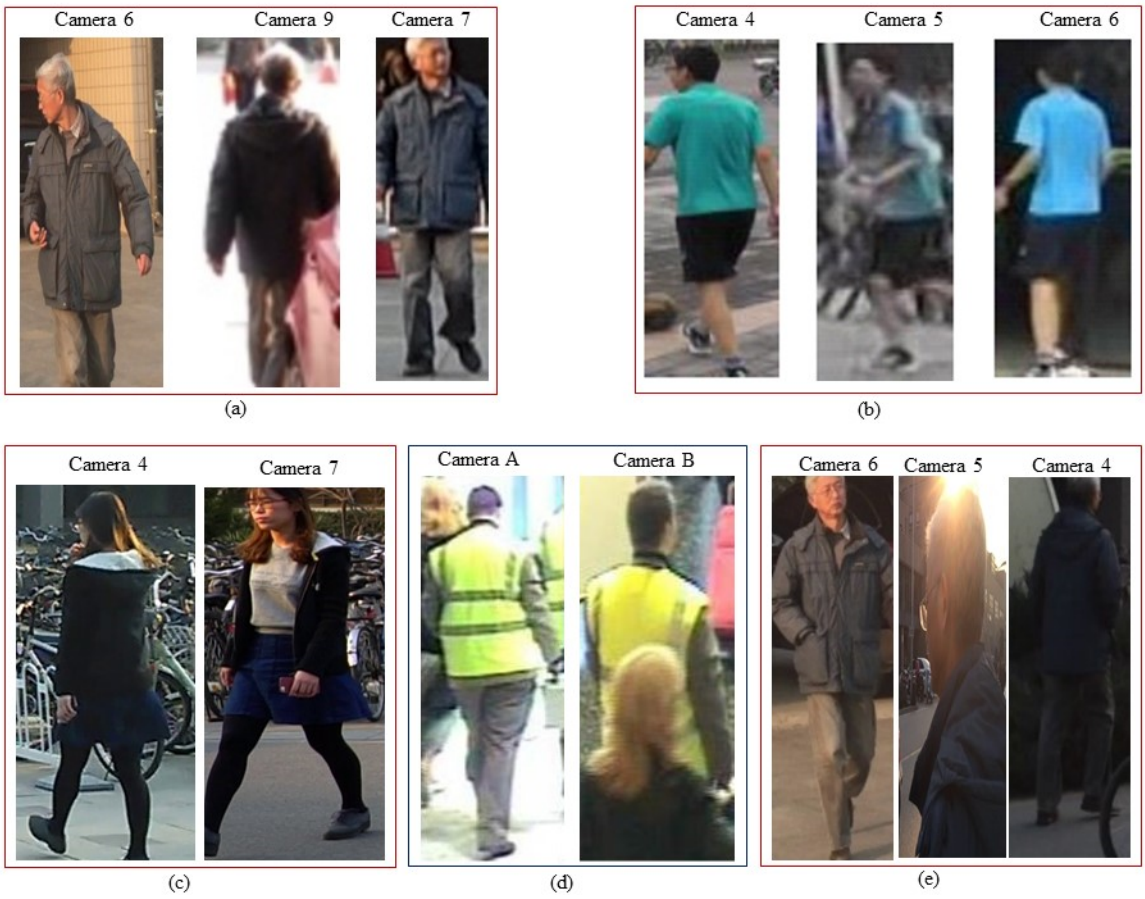

Figure 1.3: Sample images showing the challenges related to camera variations and environmental conditions in the Re-ID problem. Images are taken from the MSMT17 [4], i-LIDS [5] and Market-1501 [6] datasets.

Partial occlusion: When images are captured by the surveillance camera in a camera network, it can be partially occluded by other people or objects of the scene, or self occlusions caused by own body parts as shown in Fig. 1.3 (d).

Illumination variations: This is another biggest challenge for Re-ID. Illumination variations may occur depending on the location of the camera and environmental conditions, and it also varies in different periods of time in a day across the cameras as shown in Fig. 1.3 (e).

Background clutter: Depending on the camera location, captured images contain different objects as background scene shown in Fig. 1.3 (c). Cluttered background has a significant impact on the performance of the Re-ID because it is an obstacle to construct robust features from the captured images. When a model is trained with background cluttered images, it may suffer over-fitting problem which cause overall performance degradation.

Depending on the RGB camera or camera network, Re-ID researchers have tried to address the above challenges over the past years by proposing different computer vision and machine learning algorithms. With the advent of reliable and affordable RGB-D sensors can assist in these challenges by making use of their depth data. In this dissertation, we propose different approaches to address the extreme illumination changes and diverse cluttered background problems with RGB-D sensors for indoor applications. In addition, we also propose an effective fusion technique for multimodal data to overcome the overfitting problem that appears in the feature-level fusion method.

1.4 Objectives

Despite the recent advances, for some extreme cases (e.g. very poor lighting conditions and clothing changes), Re-ID researchers fail to address the Re-ID challenges with their proposed algorithms using traditional RGB cameras in a camera network. As modern RGB-D sensors avail us with different modalities such as illumination invariant high quality depth images, RGB images and skeleton information simultaneously as shown in Fig. 1.4, we exploit the sensor based additional information to overcome the challenges by proposing different approaches . In addition, under normal lighting conditions, we combine RGB and depth information to construct robust features to gain high re-identification accuracy.

The general objective of this thesis is to develop different models by considering different challenges for person re-identification in indoor environments using modern RGB-D sensors. The challenges are presented into the questions, which are investigated through the dissertation. The questions are:

Figure 1.4: Sample images showing different modalities such as RGB, depth and skeleton in the RobotPKU RGBD-ID [7] dataset.

1) What will happen when one or more cameras will be installed in very low lighting conditions in a camera network? (see Fig. 1.5)

2) What will happen when people change their clothes in long-term monitoring system? (see Fig. 1.6)

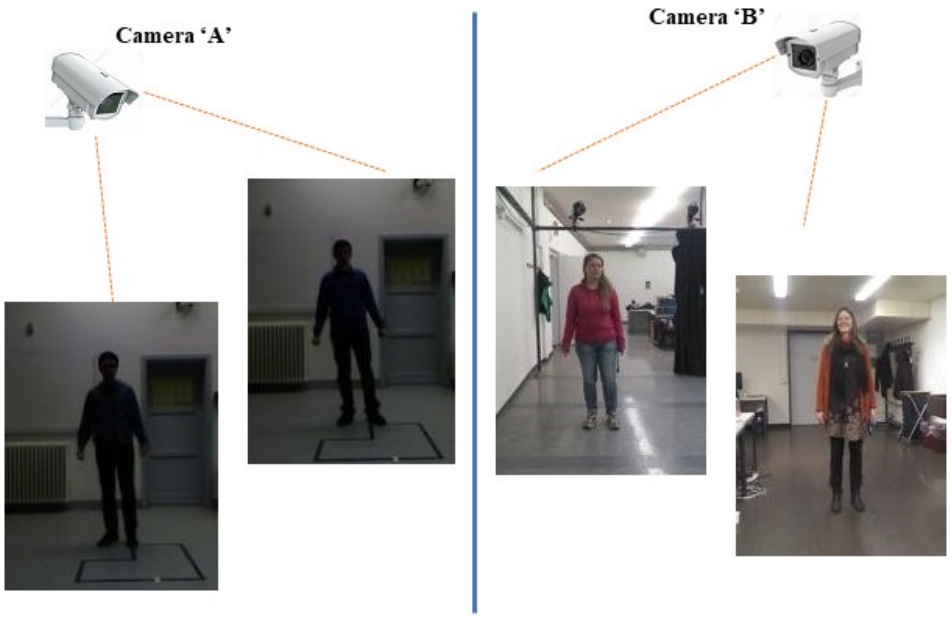

Figure 1.5: Illustration of challenges for typical re-identification under diverge lighting conditions across the cameras in a camera network.

3) Is it possible to increase the re-identification accuracy if we combinedly use RGB and Depth modalities in a system? and 4) How to combine these two modalities? (see Fig. 1.4).

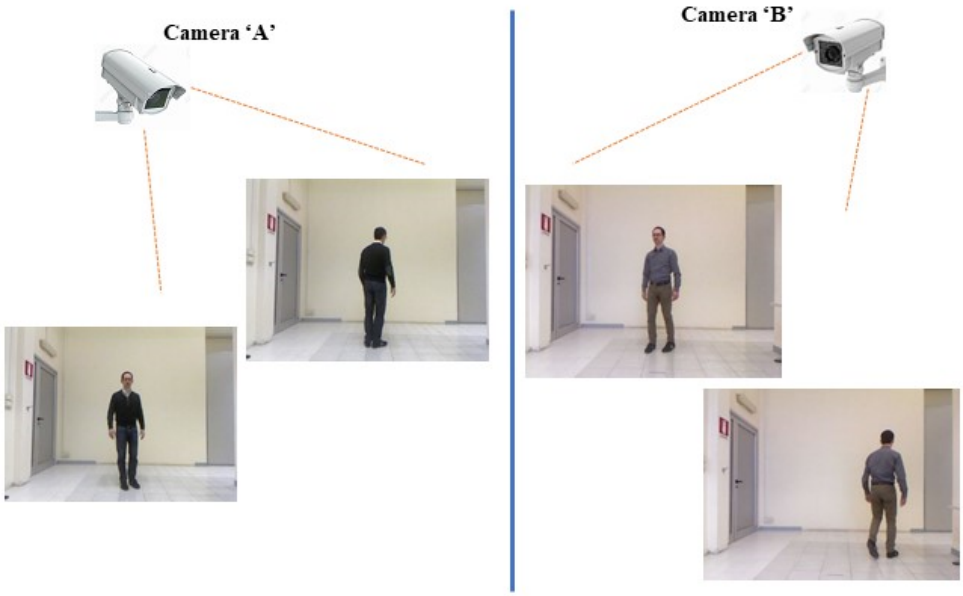

Figure 1.6: Illustration of challenges for typical re-identification while an individual changes clothes across the cameras in a camera network.

1.5 Research Contributions

As per the objectives, we pointed out two research questions (question 1 and 2), which are not possible to address using conventional Re-ID approaches because one or more camera operates in extreme low lighting conditions and remaining operates in normal lighting. Under the extreme low lighting condition, color information becomes unreachable. Moreover, when people change clothes, color becomes unreliable. In this scenario, we propose a heterogeneous camera network where depth sensor-based cameras capture illumination and color invariant depth data in poor lighting conditions, and RGB cameras capture RGB data in good lighting conditions. To that end, matching RGB images with depth images (i.e. cross-modality matching) is required, which are heterogeneous with very different visual characteristics. In this thesis, we aim at investigating this cross-modal re-identification problem, thus need to bridge the gap between these two heterogeneous modalities.

Beside the cross-modal re-identification, we also developed two deep learning Re-ID frameworks which take depth data as an additional information with RGB data to overcome background clutter problem as well as data overfitting problem for multi-modal cases to achieve higher re-identification accuracy.

The main contributions of this thesis can be summarize as follows:

- **Cross-modal Person Re-identification:** We propose a body partitioning method and HOG based feature extraction technique on both modalities, RGB and Depth domains, which extract local shape information from the images. To the best of our knowledge, this is the first attempt to extract edge gradient features on both modalities. We also exploit PCA and LDA based metric learning approach to increase re-identification accuracy. (Chapter 3)

- **Multi-modal Person Re-identification:** We propose two methods for multi-modal person re-identification. In the first method, We introduce a depth guided (DG) attention-based person re-identification framework to overcome background clutter problem. The key component of this framework is the depth-guided foreground extraction that helps the model to dynamically select the more relevant convolutional filters of the backbone CNN architecture. This leads to gain better re-identification performance. (Chapter 4)

  In the second method, we propose a re-identification approach that exploits the advantages of having multi-modal images in the form of RGB-D. In this context, we develop an effective fusion technique in dissimilarity space for 3-channel RGB and 4-channel RGB-D images to increase re-identification accuracy. (Chapter 5)

- Finally, extensive experimental evaluations demonstrate that our proposed re-identification approaches are efficient and outperform all the related state-of-the-art methods.

1.6 Thesis Overview

The remaining chapters of this thesis are structured as follows:

In Chapter 2, We briefly discuss about the state-of-the-art methods in person re-identification. The chapter is divided into several sections based on different approaches in re-identification. Most of the state of the art focuses on RGB appearance based approaches which emphasize RGB with RGB matching process (i.e. single modality matching). Others approaches (i.e. Cross-modal and Multi-modal) are also described.

In Chapter 3, We present a novel cross-modal person re-identification framework which brides the domain gap between two heterogeneous modalities (i.e. RGB and

Depth). This chapter explores how the extreme illumination and clothing changes problem were tackled by exploiting local shape information from both RGB and Depth images of an individual. A benchmark assessment is conducted by experimenting with different viewpoints in the probe and gallery samples.

In Chapter 4, We present a depth guided attention-based person re-identification method which takes into account the depth-based additional information in the form of an attention mechanism. The experimental results and discussion conclude the chapter.

In Chapter 5, we present a novel person re-identification technique that exploit the advantages of using multi-modal data for fusing in dissimilarity space, where we successfully adapt a 4-channel image input in re-identification framework. We also present the description of our collected dataset for experimental purposes. The experimental results, failure cases analysis and discussion conclude the chapter.

Finally, **Chapter 6** concludes this dissertation with a summary of the research and a discussion of the advantages and limitations of the work, as well as future perspectives.

Chapter 2

Literature Review

In the last few years, a large number of methods have been proposed for person re-identification systems, which are generally categorized into image-based and video-based Re-ID (see Fig. 2.1). In this chapter, we mainly focus image-based Re-ID and their state-of-the-art researches towards person re-identification.

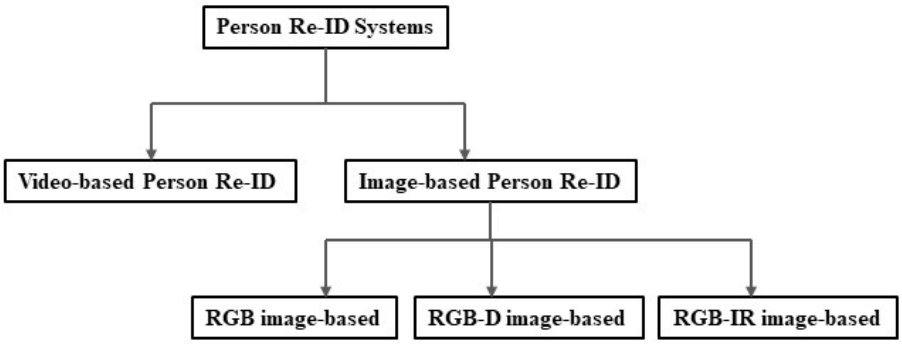

Figure 2.1: General categories of person re-identification systems.

Currently, most of the works focus on RGB image-based Re-ID. However, in some applications, RGB images are not suitable (e.g. dark environment). With the advent of modern RGB-D (e.g. Microsoft Kinect and Intel RealSense depth camera) sensors [9], RGB-D and Infrared (IR) image-based Re-ID methods have also gained increasing attention in recent years to tackle the challenges mentioned in the section 1.3.

Depending on the different modalities such as RGB, Depth and Infrared images, which are acquired by the different senors, existing image-based Re-ID works can be further subdivided into three categories as modality-aware:

- Single-modality Person Re-identification

- Cross-modality Person Re-identification
- Multi-modality Person Re-identification

2.1 Single-modality Person Re-identification

Conventional person re-identification mainly focuses on single-modality module (i.e. RGB-RGB feature matching process), where all the person images are captured by visible cameras in normal lighting conditions. For example, there are a set of probe images (RGB) and a set of gallery images (RGB) which are captured by RGB cameras, the re-identification system is usually return a ranked list of the individuals from the gallery set for each query from the probe set (see Fig. 2.2).

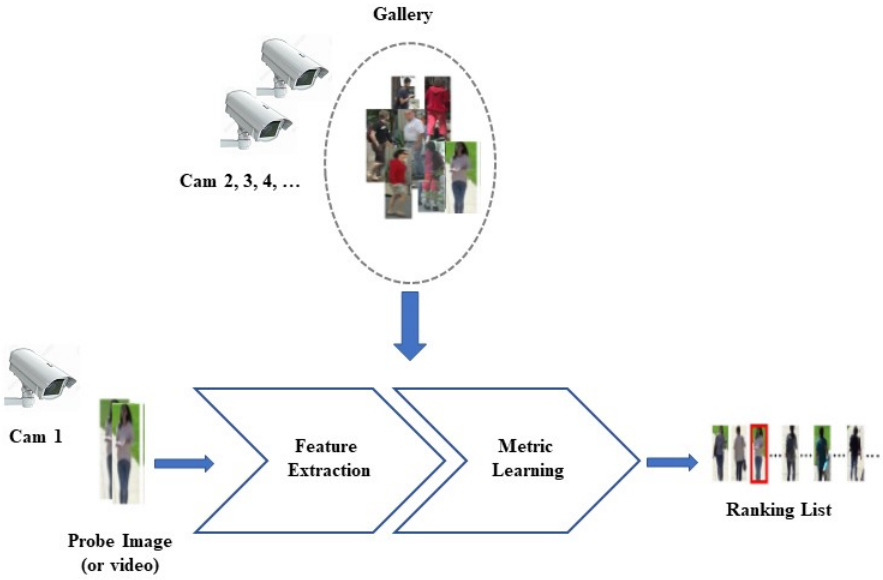

Figure 2.2: Conventional person re-identification system [15].

Existing single-modality person Re-ID approaches can be divided into three categories:

(1) Feature learning approach

(2) Metric learning approach

(3) Deep learning approach

2.1.1 Feature Learning approach

Feature learning approaches focus on learning a discriminative and robust feature representation to design a powerful descriptor (or signature) for each individual regardless of the scene [10, 11, 12, 13, 14]. In [10], the authors propose a Covariance descriptor based bio-inspired features for person re-identification. At first, they extract Biologically Inspired Features (BIF) from an individual image, and then the Covariance descriptor is used to compute the similarity of BIF features taken at neighboring scales where covariance descriptors can capture shape, location and color information. In [11], Bhuiyan et al. propose a re-identification method by segmenting the pedestrian images into meaningful parts, then extract features from such parts as well as from the whole body and finally, perform a salience analysis based on regression coefficients. Liao et al. [12] propose an efficient feature representation called Local Maximal Occureence (LOMO), and a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA) for person re-identification. The LOMO feature analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation against viewpoint changes. In [13], the authors propose a salient color names based color descriptor (SCNCD) for person re-identification. Based on SCNCD, color distributions over color names in different color spaces are obtained and fused to generate a feature representation. Zheng et al. [14] propose an unsupervised Bag-of-Words (BoW) descriptor for scalable person re-identification. In the BoW model, local features are quantized to visual words using a pretrained codebook.

2.1.2 Metric Learning approach

In metric learning approaches usually aim at learning a discriminative distance measurements to measure the similarity on top of the learned features [16, 17, 18, 19]. In [16], the authors use an efficient asymmetric metric learning for person re-identification where they consider a positive semi-definite (PSD) constraint to provide a useful regularization to smooth the solution of the metric, and hence the learned metric is more robust than without the PSD constraint. Wang et al. [17] propose a cross-scenario transfer person re-identification system which maximize the cross-task data discrepancy on the shared components during asymmetric multitask learning (MTL), along with maximizing the local inter-class variation and minimizing local intra-class variation on all tasks.. The authors in [18], formulate an asymmetric distance model for learning camera-specific projections for person re-identification where they transform

the unmatched features of each view into a common space, and then discriminative features across view space are extracted. In [19], Zheng et al. formulate person re-identification as relative distance comparison (RDC) learning problem in order to learn the optimal similarity measure between a pair of person images.

2.1.3 Deep Learning approach

In recent years, deep learning approaches for Re-ID have received substantial attention due to their powerful deep features, which enable to obtain good performance compared to hand-crafted features, specially for large dataset. The general architecture for deep learning person Re-ID system is shown in Fig. 2.3.

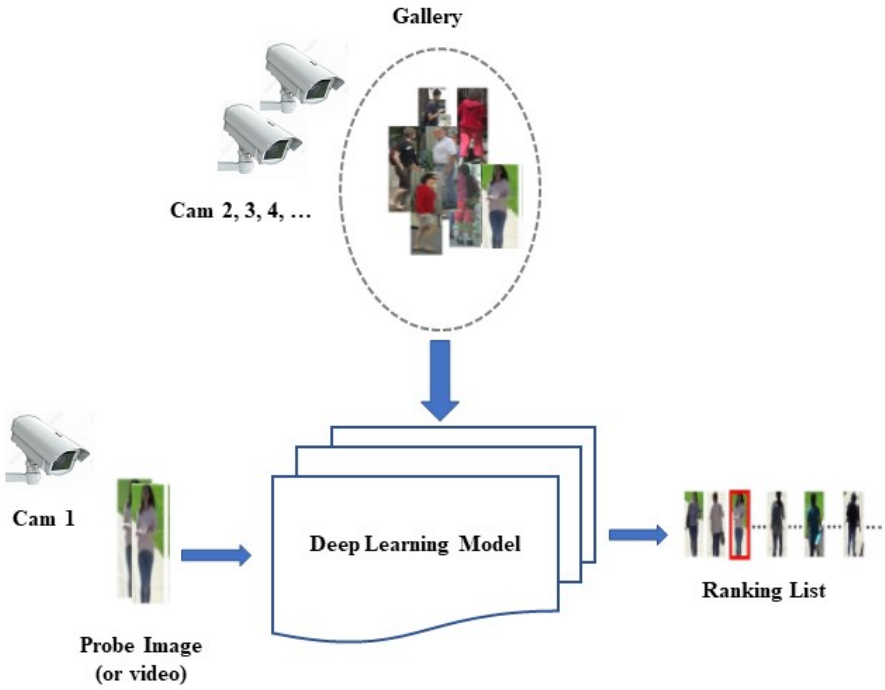

Figure 2.3: Deep learning person re-identification system [15].

The main idea of using deep learning architecture of person re-identification comes from Siamese CNN with either two or three branches for pairwise verification loss [20, 21, 22, 23, 24, 25] or triplet loss [26, 27, 28] respectively, or combination of both [29]. In [21], the authors present a new deep convolutional architecture for person Re-ID by designing two special layers for capturing relationships between two views: a cross-input neighborhood differences layer, and a subsequent layer that summarizes these differences. Some approaches [22, 30] fuse features from different body parts

with a multi-scale CNN structure. To obtain superior accuracy, some re-identification methods [26, 28, 31] use the pre-trained or different variants of pre-trained models(e.g. ResNet [32], GoogleNet [33]). Transfer learning is another form of deep learning architecture which is successfully applied in person Re-ID approaches [24, 34, 35], when the distribution of the training data from the source domain is different from that of the target domain.

All the above mentioned approaches (i.e. Feature learning, Metric Learning and Deep Learning) have been effective but in the situations where people may change their clothing in long-term monitoring or in dark environments, these RGB-based appearance features tend to fail.

2.2 Cross-modality Person Re-identification

Different from the aforementioned single-modality person Re-ID , cross-modality person Re-ID aims to match quires of one modality against a gallery set of another modality, such as RGB-Depth Re-ID and RGB-Infrared(IR) Re-ID. Fig. 2.4 illustrates a typical cross-modal person re-identification system.

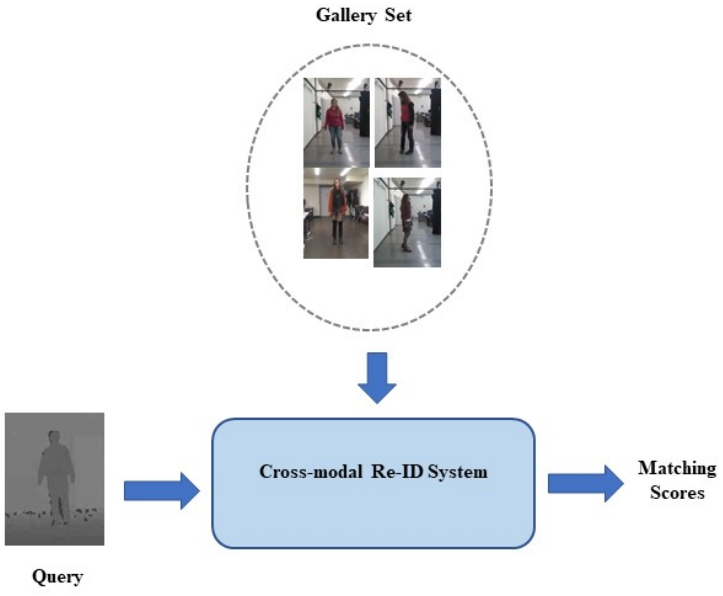

Figure 2.4: A typical cross-modal person re-identification system based on RGB (gallery set) and depth (query) modalities.

Beside the well known single-modality person Re-ID, very few works have been

investigated for RGB-Depth [36, 37] cross-modal person Re-ID. In [36], Zhuo et al. perform cross-modal person re-identification between depth and RGB on heterogeneous camera networks. They propose a dictionary learning based method to encode different-modality body shape features such as edge gradient feature and Eigen-depth feture which are extracted from the RGB and depth domain respectively. In [37], the authors propose a cross-modal distillation network for robust person re-identification.

Recently, some works [38, 39, 40, 41, 42, 43, 44, 45] perform RGB-IR cross-modal person Re-ID. Lu at al. [38] propose a cross-modality person re-identification wih shared-specific feature transfer. Dai et al. [39] design a cutting-edge generative adversarial training based discriminator to learn discriminative feature representation from RGB and infrared modalities. In [41, 42], Ye at al. advance a two-stream based model and bi-directional top-ranking loss function for the shared feature embedding. In [43], the authors propose to generate cross-modality paired-images and perform both global set-level and fine-grained instance-level alignments. Wang et al. [45] introduce dual-level discrepancy reduction learning based on a bi-directional cycle GAN to reduce the gap between RGB and depth modalities. Wang et al. [44] construct a novel GAN model with the joint pixel-level and feature-level constraint, which achieve the state-of-the-art performance.

In this thesis, we consider RGB-Depth cross-modal person re-identification where we extract local shape information from partitioned regions of the body of an individual for the RGB and depth domains on the heterogeneous camera networks, and also exploit PCA and LDA based metric learning approach to increase re-identification accuracy.

2.3 Multi-modality Person Re-idetification

As most of the current Re-ID methods focus on matching individuals based on traditional RGB cameras, some constraints such as illumination and clothing changes cannot properly be addressed using RGB cameras. After the arrival of RGB-D sensors, Re-ID researchers took advantage of other modalities such as depth and skeleton information to address the above-mentioned problems, and to increase Re-ID accuracy as well. In this section, we overview the RGB-D sensor-based person re-identification methods which are most relevant to our thesis work.

In the RGB-D based Re-ID literature, some re-identification methods have been proposed based on depth images, point clouds and anthropometric measurement to

solve the problems of changing clothing (i.e. for long-term re-identification) and extreme illumination [46, 47, 48, 49, 50, 51, 52, 53]. Though RGB-D sensors can capture RGB, depth and skeleton information simultaneously, however when people appear in excessive lighting environments or change clothes, in this case some authors consider only depth-based person Re-ID [51, 52] approaches to solve such constraints. In [51], Haque et al. propose a recurrent attention model for depth-video-based person identification, in which a 3D RAM model is for still 3D point clouds and a 4D RAM model is for 3D point cloud sequences. However, Haque's method is not suitable for solving the person re-identification problem under the setting when there is no overlap between identities in training and testing. In [52], the authors propose an approach for long-term person re-identification by using depth videos where they develop a sparse canonical correlation analysis using a local third-order tensor model to perform multi-level person Re-ID.

In some works, authors propose skeleton-based anthropometric measures for person re-identification [47, 49, 50]. Barbosa et al. [50] use skeleton-based features based on anthropometric measurements of the Euclidean distance between selected body parts such as legs, arms and the overall height, and geodesic distances on the body surface. The geodesic distances are computed from a predefined set of joints (e.g. from the torso to the right hip). In [49], the authors propose two kinds of descriptors where the first descriptor contains anthropometric measures computed from body joint points and the other descriptor contains a point cloud model of the human body. In [47], Munaro et al. modified the work proposed in [50] by combining Point Cloud Matching (PCM) and skeleton-based features. Although these works use depth-based point clouds and skeleton information to tackle the pose variations of a person, they do not perform any feature-level fusion or score-level fusion techniques.

Apart from them, some works [46, 48, 53] propose two different types of features extracted from a given depth image and skeleton joint points accordingly, and then finally fused by score-level fusion to gain high re-identification accuracy. Wu et al. [46] propose to exploit depth information to provide a depth voxel covariance descriptor and rotation invariant depth shape descriptor called Eigen-depth feature. To enrich the depth shape descriptor, they also employ a skeleton-based feature as complementary physical information. In this work, they calculate the Euclidean distance between skeleton-based features, and the geodesic distance be-tween the corresponding within-voxel covariance matrices and between-voxel covariance matrices. Finally, they measure the similarity of two subjects by summing both distances. In [48], Imani

et al. extract three types of histogram features (Local Binary Patterns (LBP), Local Derivative Patterns (LDP) and Local Tetra Patterns (LTrP)) from depth images where at first depth images are divided into three regions of head, torso and legs using skeleton data. Then these histogram features are fused with anthropometric features (where anthropometric features are calculated from skeleton joint points) using score-level fusion. In [53], the authors introduce two novel features: histogram of the edge weight (HEW) and histogram of the node strength (HNS) where these features fit both single-shot and multi-shot person Re-ID. Then these features are combined with skeleton features using score-level fusion.

Some authors have proposed some conventional Re-ID methods to combine RGB appearance cues with other modalities, such as depth, thermal data, gait and anthropometric measures [54, 55, 56, 57, 58]. In [54], the authors propose Skeleton Standard posture (SSP) and color descriptors from RGB-D data (color point clouds). A partition grid is computed to extract color-based features through the SSP. Then, the extracted features from the database are re-projected using the partition grid under investigation. Finally, these extracted features are used to determine people's differences. Pala et al. [55] fuse clothing appearance descriptors with anthropometric measures extracted from depth data to improve re-identification accuracy. They also propose a dissimilarity-based framework for building and fusing multi-modal descriptors of pedestrian images, which is an alternative of score-level fusion. In [56], Mogelmose et al. propose a tri-modal re-identification method to combine RGB, depth and thermal features. The modalities are combined in a late fusion strategy, which is able to predict a new subject in the scene as well as to recognize previous subjects based on a combined rule cost. Kawai et al. [57] introduce a view-dependent score-level fusion method to combine color and gait features. In [58], the authors propose an online re-identification method based on metric model update for robotic applications. In this method, each person is described by appearance and geometric features using skeleton information. Then a fusion technique named Feature Funnel Model (FFM) is proposed to fuse multi-modal features effectively.

Recently, a few works [59, 60, 61, 62] based on deep learning methods have been proposed for RGB-D multi-modal person re-identification. In [59], the authors propose a multi-modal uniform deep learning method to extract the RGB appearance feature and anthropometric features from processed depth images. The proposed method uses two CNNs for separately analyzing the depth and RGB images. Afterwards, they design a multi-modal fusion layer to combine these features extracted from both depth images and RGB images with a uniform latent variable. In [60], Ren

et al. propose a uniform and variational deep learning method for RGB-D object recognition and person re-identification. This method extracts the depth feature and the appearance feature from the depth and RGB images with two CNNs respectively. The depth feature and appearance feature are then combined with a variational auto-encoder at the top layer of their proposed deep network. Lejbolle et al. [61] propose a multi-modal CNN which is trained using both depth and RGB modalities to provide a fused feature. Later the authors improve their approach with a multi-modal attention network [62], in which they add an attention module to extract local and discriminative features that are fused with globally extracted features.

In this thesis, we consider RGB and depth modalities, and propose two new deep learning Re-ID frameworks for multi-modal cases. The key component of our first deep learning Re-ID framework is the depth-guided foreground extraction by removing background clutter of an individual image, which helps the model to dynamically select the more relevant convolutional filters of the the backbone CNN architecture. In our second method, we use two individually trained models for RGB-D person re-identification, where models are trained using 3-channel RGB and 4-channel RGB-D images accordingly. Then the dissimilarity score is calculated using feature embeddings extracted from both trained models and finally fuse both scores in dissimilarity space. As some of the above state-of-the-art Re-id approaches use multi-modal fusion in feature space, which may cause overfitting problem due to the noisy/heterogeneous feature. Different from them, we utilize the ensemble of RGB and RGB-D based trained models in dissimilarity space which assists to overcome the overfitting problem due to noise.

There are some state-of-the-art methods in the Re-ID task [63, 64, 65, 66] to handle the background clutter problem in RGB images using whole body attention and part-based attention mechanisms. The first key ingredient of these approaches is human body mask generation which is very costly in computation. These methods obtain human body masks using different deep learning based image segmentation models such as FCN [67], Mask R-CNN [68], JPPNet [69], and Dense Pose [70]. All these sate-of-the-art approaches heavily depend on very complex dedicated attention-based architectures which involve large computational costs. For this reason, it is difficult to deploy for them in real time scenario. In contrast to the above works, we introduce a depth guided attention mechanism for person re-identification with less computational effort.

Chapter 3

Cross-modal Person Re-identification using Local Shape Information

3.1 Introduction

Recent Re-ID research has mainly focused on RGB-RGB matching, which is the most common scenario where there is only a single-modality. However, RGB based Re-ID systems have limitations in surveillance when lighting is either very poor, since RGB based cameras cannot capture sufficient information in dark environments. Moreover, when people change their clothes in the long-term monitoring system, color becomes unreliable (see Fig. 3.1). In comparison to RGB cameras, depth cameras can capture video even in low lighting conditions. So, it is possible to extract depth information and the body skeleton using depth cameras [46] (e.g., Microsoft Kinect) in dark environments. The Kinect sensor in particular, can capture the depth information of each pixel by using an infrared sensor, regardless of the pedestrian's color appearance and illumination in indoor environments (see Fig. 3.2).

Most existing works in person Re-ID emphasize on either RGB camera networks or depth camera networks [46, 51]. Some recent works utilize RGB-Infrared heterogeneous camera network [38, 39, 40, 41, 42]. While less sensitive than RGB, infrared cameras can still be affected by illumination changes from real-world environments, as well as temperature changes in the environment. Very few works use RGB-depth heterogeneous camera network for RGB-Depth cross-modality matching [36, 37]. In [37], the authors propose a cross-modal distillation network for person re-identification across RGB and depth sensors while Zhuo et al. [36] proposes a dictionary learning based method on heterogeneous camera networks that contain RGB and depth im-

Figure 3.1: Illustration of change of light and clothes across different cameras in different times and locations. (a) Same person in different lighting conditions. (b) Same person in different times of the day with different clothes

ages. Specifically, the authors in [36] proposed two kinds of edge gradient features for RGB images, which are the classic Histogram of Oriented Gradient (HOG) [71] and Scale Invariant Ternary Patterns (SILTP) [12]. Both of them can describe the body's shape coarsely. For depth images, they extract Eigen-depth features from 3D point clouds of segmented torso and head parts only.

In this work, we propose a body partitioning method and HOG based feature extraction technique on both modalities because it captures edge or gradient structures which represent local shapes in scenes. In [36], they extract features only from segmented regions (head and torso part) from the depth domain. However, to the best of our knowledge, our work is the first attempt to extract edge gradient features on both RGB and depth modalities at the same time. To learn discriminant features, we first apply Principle Component Analysis (PCA) for dimensionality reduction, then we exploit the beneficial properties of Linear Discriminant Analysis (LDA) within the PCA subspace to find the low intra-class variation and high inter-class variation of the data. This allows us to gain good performance than the existing methods for the task of person re-identification on heterogeneous camera networks.

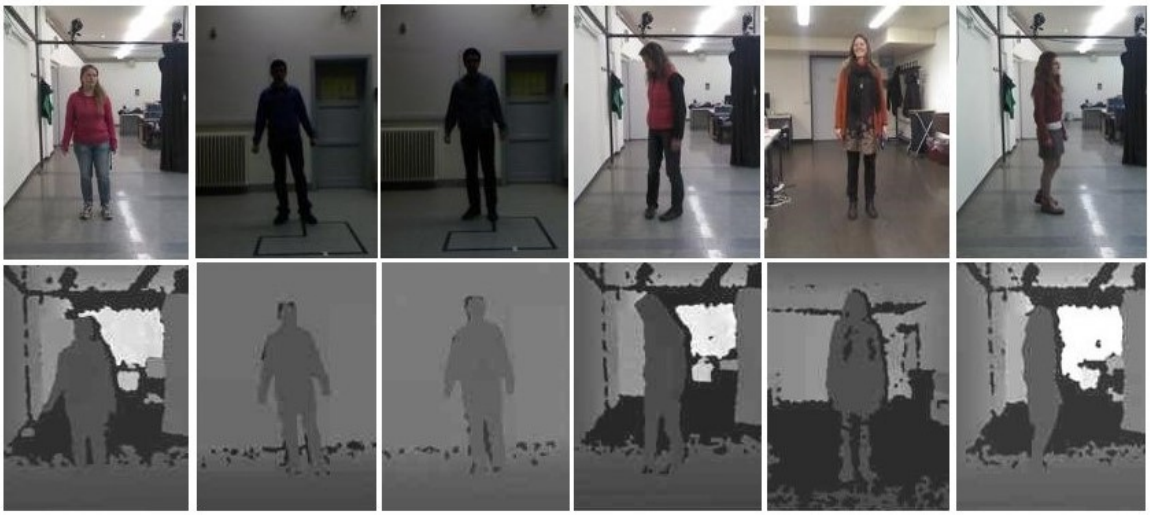

Figure 3.2: Examples of RGB and depth images captured in indoor environments. In Row 1, columns 1, 4, 5 and 6 show the RGB images in good illumination conditions, with columns 2 and 3 in poor illumination conditions accordingly. Row 2 shows the depth images of all RGB images.

We tested our methods on two publicly available datasets, the BIWI RGBD-ID [49] and IAS-Lab RGBD-ID [47]. Our contributions can be summarized as follows:

1. We propose a body partitioning method and HOG based feature extraction technique on both modalities, RGB and Depth domains, which extract local shape information from the image. To the best of our knowledge, this is the first attempt to extract edge gradient features on both modalities.

2. We exploit PCA and LDA based metric learning approach to increase re-identification accuracy.

3. Extensive experiments show the effectiveness of the proposed method over two RGB-D benchmark re-identification datasets.

3.2 Methodology

Our re-identification approach has three distinct phases: (1) Feature extraction, (2) Metric learning, and (3) Feature matching. The overall system is illustrated in Fig. 3.3.

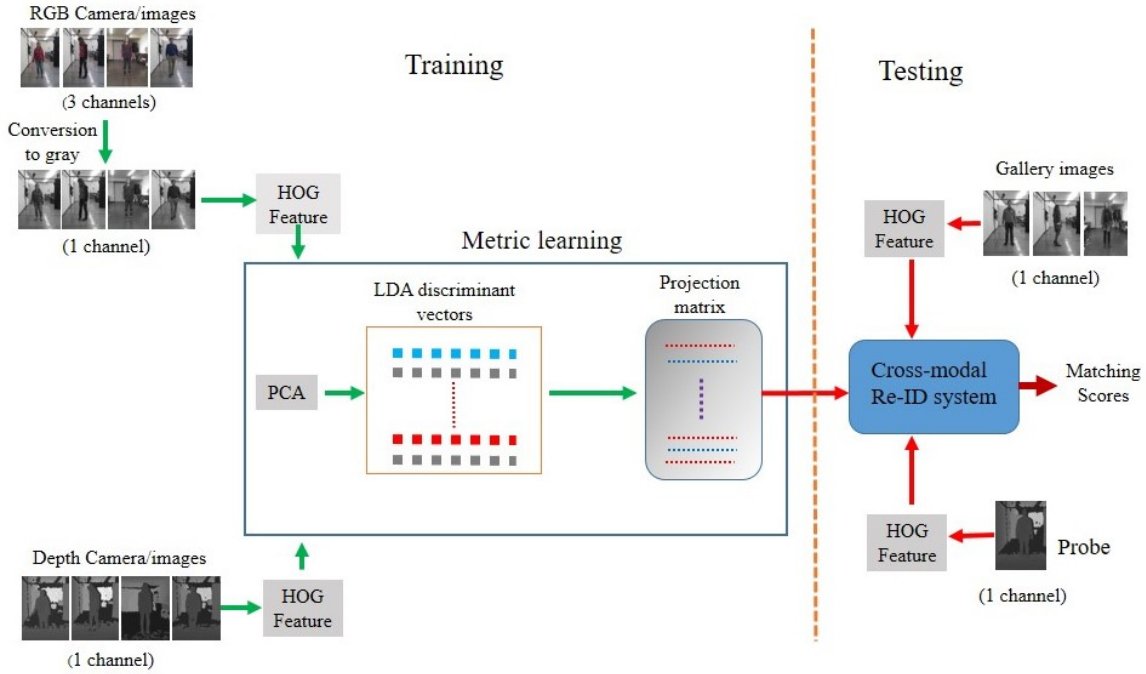

Figure 3.3: Overview of our proposed approach. In the training stage, labeled image pairs from RGB and depth cameras are used to jointly learn the discriminative features by LDA. After dimensionality reduction, the projected features are matched by using Euclidean distance in the testing stage.

3.2.1 Feature extraction

In this section, we give the details of feature extraction using HOG [71], which extracts features from both camera images. HOG has been widely accepted as one of the best features to capture edges or local shape information. Though HOG can extract features from a true color (RGB and LAB color spaces) or grayscale images, we find extracting features from grayscale images works best in our RGB-depth setup. According to our proposed method (see Fig. 3.3), RGB images are captured by an RGB camera and depth images captured by a depth camera (e.g. Kinect or Intel RealSense Depth camera) on the heterogeneous camera network. To facilitate cross-modal learning, our aim is to first make the images from the RGB and depth domains as similar as possible. Since RGB images have three color channels, we need to convert it to a single channel for convenience because the Kinect sensor depth images are 16-bit depth monochrome images with 65,536 levels of sensitivity.

In this work, we divide a person image into six horizontal stripes (see Fig. 3.4). This is a generic human body partitioning method that is widely used in existing methods [72, 73] to capture distinct areas of interest. For HOG features, each strip

is further divided into 2×2 blocks of 8×8 pixel cells with 50% overlapping blocks, and each cell contains 9 orientation bins. Each strip returns the features as a 1-by-v1 vector. Finally, the feature vectors of all 6 strips are concatenated to construct a final feature vector of 1-by-d, where d=v1 + ... +v6, is inside an image window. Since the histograms are computed for regions of a given size within a window, HOG is robust to some location variability of body parts. HOG is also invariant to rotations smaller than the orientation bin size.

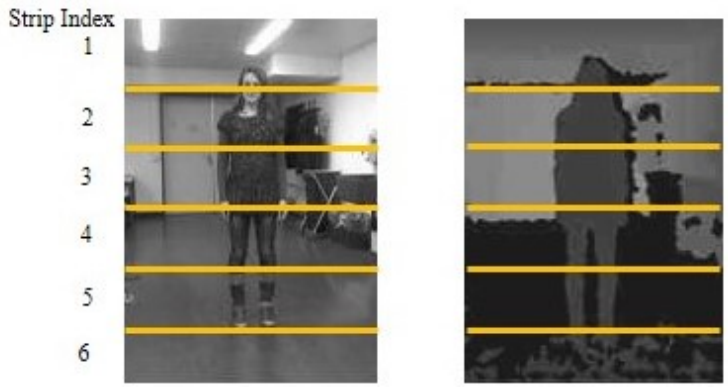

Figure 3.4: A spatial representation of human body is used to capture visually distinct areas of interest. The representation employs six equal-sized horizontal strips in order to capture approximately the head, upper and lower torso and upper and lower legs.

3.2.2 Metric learning

In the metric learning approach, we first extract features for each image, and then learn a metric with which the training data have strong inter-class differences and intra-class similarities. In such a case, we employ linear discriminant analysis (LDA) to determine a set of projection vectors maximizing the between-class scatter matrix (S_b) while minimizing the within-class scatter matrix (S_w) in the projective space. However, LDA often suffers from issues such as small sample size and high dimensionality (so there are too many variables). When there are not enough training samples and/or the dimensionality is too high, (S_w) may become singular, and it is difficult to compute the LDA vectors. In our work, we use a two-stage approach PCA+LDA [74] to address this problem. First, we reduce the feature dimensionality using Principal Component Analysis (PCA), and then LDA is applied on the reduced PCA subspace, in which (S_w) is non-singular.

LDA tries to find the projection matrix W maximizing the ratio of the determinant of S_b to S_w,

$$W = \arg\max \left| \frac{W^T S_b W}{W^T S_w W} \right| \tag{3.1}$$

Consider that the training set contains C classes which are taken from the RGB camera and depth camera, and each class X_i has n_i samples. S_b and S_w are defined as,

$$S_b = \sum_{i=1}^{C} n_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{3.2}$$

$$S_w = \sum_{i=1}^{C} \sum_{\bar{x}_k \in X_i} (\bar{x}_k - \mu_i)(\bar{x}_k - \mu_i)^T \tag{3.3}$$

where μ is the mean of all data, μ_i is the mean for the class X_i, and $\bar{x}_k$ is the sample belonging to class X_i. W can be computed from the eigenvectors of $S_w^{-1} S_b$ [75]. The eigenvectors corresponding to the first m largest eigenvalues are used to construct the projection matrix.

3.2.3 Feature matching/classification

After obtaining the projection matrix, we aim to recognize a certain person on heterogeneous camera network. The goal of our cross-modal person re-identification system now is to find a person image that has been selected in the depth camera (probe image) in all images from the RGB camera (gallery images). This is obtained by calculating the Euclidean distances between the probe image and all gallery images using the learned metric, and returning those gallery images with the smallest distances as potential feature matches.

3.3 Experiments

In this section, we evaluate the performance of our approach by performing experiments on two RGB-D person re-identification datasets BIWI RGBD-ID [49] and IAS-Lab RGBD-ID [47] recorded by Microsoft Kinect cameras. Both datasets target long-term people re-identification from RGB-D cameras. In our work, besides the HOG based feature extraction technique, we also experimented on two well-known local shape descriptors SILTP [12] and LBP [76, 77] for both datasets. The SILTP descriptor is an improved operator over LBP [12].

3.3.1 Datasets

BIWI RGBD-ID [49]. This dataset has three groups of sequences, namely "Training", "Still" and "Walking", each of which contains groups of 50, 28 and 28 people respectively with different clothes, and collected on different days and in a different scenes. Some sample images are shown in Fig. 3.5, which are taken from "Training" and "Still" groups. Each person is associated with about 300 sequence of frames of depth images, RGB images and skeletons. The BIWI dataset consists of RGB images with a resolution of 1280×960 and depth images with a resolution of 640×480.

Figure 3.5: Example of the RGB and their corresponding depth images of the same person with different clothes, and captured on different days and in different indoor locations.

IAS-Lab RGBD-ID [47]. In the IAS-Lab RGBD-ID dataset, there are 11 different people. This dataset contains three groups of sequences "Training", "TestingA" and "TestingB", and each person performs out-of-plane rotations on himself and walks in the recordings. There are about 500 frames of depth images, RGB images and skeletons for each person. The first (Training) and second (TestingA) sequences were acquired when same person was wearing different clothes (see Fig. 3.6), while the third one (TestingB) was collected in a different room, but with the same clothing as in the first group (Training). Some sequences in "TestingA" and "TestingB" were recorded under low lighting (see Fig. 3.7). The IAS-Lab dataset consists of RGB images with a resolution of 640×480 and depth images with a resolution of 640×480.

Figure 3.6: Example of the RGB and their corresponding depth images of the same person with different clothes.

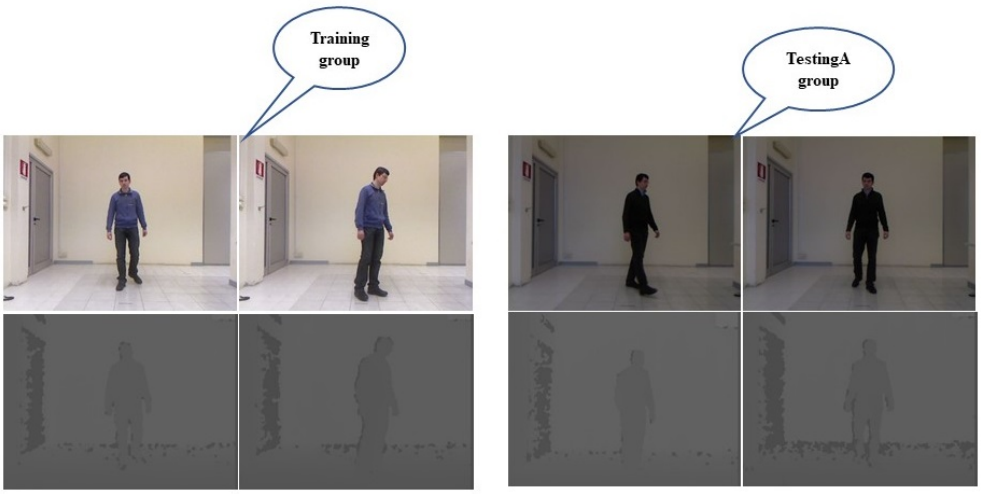

Figure 3.7: Example of the RGB and their corresponding depth images of the same person with different lighting conditions.

Data Pre-processing. As depth images are single channel, so we convert all RGB images to gray scale images. Before HOG feature extraction, all depth and RGB images are resized to 256×192 to maintain the original aspect ratio of the images, which retain edge gradient shape without distortions.

3.3.2 Evaluation Metrics

We show the results in terms of recognition rate as a cumulative matching characteristic (CMC) curve and rank-k accuracy, which are common practice in the Re-ID literature [46]. Rank-k accuracy is the cumulative recognition rate of correct matches at rank k. The CMC curve represents the cumulative recognition rates at all ranks. The evaluation is repeated 10 times and the average results are reported. For quantitative evaluation, the average rank 1, 5 and 10 accuracy performance measures are reported. In this work, all results are reported using the single-shot strategy, where one image per sample is randomly selected as the gallery. Even though multiple images (which refer as multi-shot) can be used as query and gallery set to increase the re-identification accuracy, it computational cost is high over single-shot.

3.3.3 Compared Methods

To evaluate the effectiveness of our approach, we compare our method with a recently proposed cross-modal re-identification approach on a heterogeneous camera network [36, 37]. In [36], the authors performed the Re-ID task across the depth and RGB modalities and proposed a dictionary learning based method to encode different-modality body shape features including an edge gradient feature and the Eigen-depth feature for the BIWI RGBD-ID and RGBD-ID datasets. In [37], a deep neural networks is proposed for cross-modal person re-identification between RGB and depth modalities. In our work, we use PCA and LDA based metric learning method for edge gradient feature extraction on both modalities. Besides HOG features, we also investigate two local body shape descriptors including SILTP and LBP on our proposed approached. These feature descriptors are extracted using same algorithm for both modalities. LBP has a nice invariant property under monotonic gray-scale transforms, but it is not robust to image noise. SILTP improves on LBP by introducing a scale invariant local comparison tolerance and robustness to image noise [12].

3.3.4 Evaluation on BIWI RGBD-ID

We use the complete "Training" and "Still" groups in our experiment, hence there are 78 video sequences (samples) in total. As in [36] same person with different clothing is considered as a separate instance. we randomly select five frames each from the RGB and depth video sequences for each sample. By convention, we randomly choose about half of the samples, 40 pedestrians for training and the remaining for testing.

Each experiment is carried out in two cases. For the first case, we select RGB images as the gallery and depth images as the probe, and in the second case, we use depth images as the gallery and RGB images as the probe. Table 3.1 reports the results with single-shot setting and compare with other existing methods [36, 37], where in [36] no detailed information on the evaluation procedure is given. In [36, 37], public codes are not available. It is also hard and time consuming to implement the all methods. Therefore, we couldn't verify the results under this setting.

Table 3.1: Average accuracy of the existing methods and our proposed approach for different scenarios on the BIWI dataset.

Approaches	Gallery-RGB, Probe-Depth			Gallery-Depth, Probe-RGB		
	rank-1 (%)	rank-5 (%)	rank-10 (%)	rank-1 (%)	rank-5 (%)	rank-10 (%)
Eigen-depth HOG, CCA [36]	6.31	27.63	40.79	6.31	24.21	40.79
Eigen-depth SILTP, CCA [36]	6.58	27.37	45.00	8.42	26.32	41.58
Eigen-depth HOG, LSSCDL [36]	7.11	28.42	41.32	8.42	27.11	46.05
Eigen-depth SILTP, LSSCDL [36]	7.37	29.47	50.26	9.47	24.21	40.26
Eigen-depth SILTP, Dictionary learning [36]	9.21	26.32	46.05	12.11	26.32	41.58
Eigen-depth HOG, Dictionary learning [36]	11.32	30.26	48.16	11.84	28.42	44.47
Cross-modal distillation network [37]	29.23	70.50	88.13	26.85	65.88	84.13
LBP, PCA+LDA metric learning **(Ours)**	35.01	82.51	95.08	34.30	82.53	95.21
SILTP, PCA+LDA metric learning **(Ours)**	36.89	84.20	96.52	36.14	83.34	95.21
HOG, PCA+LDA metric learning **(Ours)**	41.43	82.51	94.36	36.52	79.73	92.38

Table 3.1 shows that our approach outperforms all the existing methods [36, 37]. In [36], the authors also compared their method with Least Square Semi-Coupled Dictionary Learning (LSSCDL) [78] and Canonical Correlation Analysis (CCA) [79].

In [37], though the authors proposed a deep neural network to transfer knowledge from one modality to a second modality to solve the re-identification task across the two modalities (RGB and depth), they failed to achieve good result because of the intrinsic nature of data. In the results, we also see that when LBP and SILTP features are extracted from both modalities and we apply our metric learning approach, then results also outperform the method proposed by [36, 37] on the heterogeneous camera network. Our method achieves 41.43%, 36.89% and 35.01% rank-1 accuracy for HOG, SILTP and LBP features, respectively when we select RGB as gallery and depth as probe. However, when we choose depth as gallery and RGB as probe, we obtain 36.52%, 36.14% and 34.30% rank-1 accuracy, which are slightly lower than previous settings. The average results of the three local shape descriptors with our metric learning approach is shown in Fig. 3.8 using a CMC curve over 10 trials.

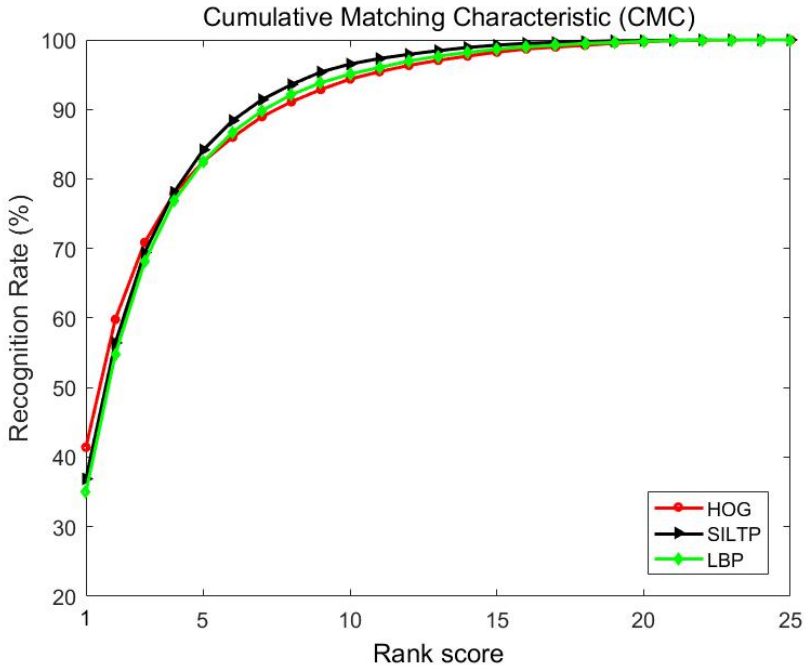

Figure 3.8: Performance on BIWI RGBD-ID (single-shot) dataset for three local shape descriptors with our approach, where we set Gallery-RGB and Probe-Depth images.

3.3.5 Evaluation on IAS-Lab RGBD-ID

On this dataset, the evaluation also follows the same settings as with the BIWI dataset with one exception. In this experiment, we randomly select ten frames from

the RGB and depth images to avoid singularity problem with LDA. We use the complete "Training" and "TestingA" groups in our experiment, hence there are 22 samples in total. By convention, we randomly choose exact half of the samples, 11 pedestrians for training and the remaining for testing. The average rank-1, rank-5 and rank-10 accuracies over 10 trials of evaluation are reported in Table 3.2. The performance of the tested methods is shown in Fig. 3.9 and 3.10 using a CMC curve over 10 trials.

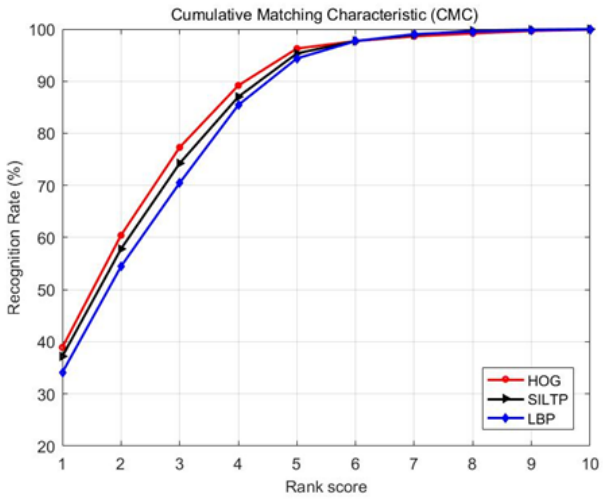

Figure 3.9: Performance on IAS-Lab RGBD-ID (single-shot) dataset for three local shape descriptors with our approach, where we set Gallery-RGB and Probe-Depth images.

Table 3.2: Average accuracy of our proposed approach for different scenarios on the IAS-Lab dataset.

Approaches	Gallery-RGB, Probe-Depth			Gallery-Depth, Probe-RGB		
	rank-1 (%)	rank-5 (%)	rank-10 (%)	rank-1 (%)	rank-5 (%)	rank-10 (%)
LBP, PCA+LDA metric learning **(Ours)**	34.11	94.38	99.94	33.71	95.45	99.74
SILTP, PCA+LDA metric learning **(Ours)**	37.20	95.33	100	35.24	96.04	99.78
HOG, PCA+LDA metric learning **(Ours)**	38.93	96.28	99.89	38.21	96.80	99.99

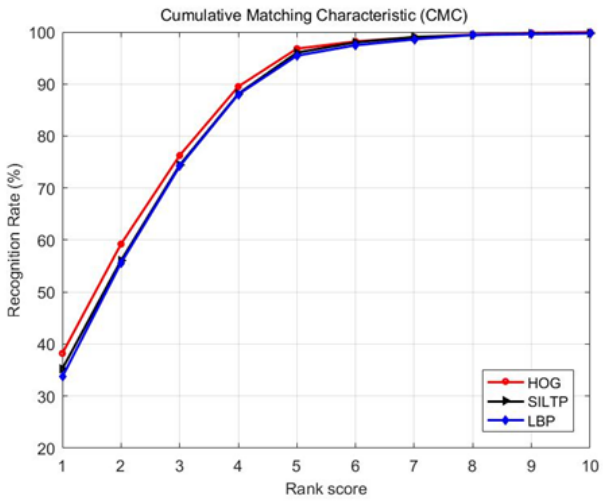

Figure 3.10: Performance on IAS-Lab RGBD-ID (single-shot) dataset for three local shape descriptors with our approach, where we set Gallery-Depth and Probe-RGB images.

As observable, we obtain better result on both datasets (BIWI and IAS-Lab) when we select RGB images as gallery and depth image as probe. This implies that our method is consistent and effective for this setting.

3.4 Conclusion

In this paper, we have presented a cross-modal re-identification system for RGB and depth heterogeneous camera networks. This is in contrast to most existing camera networks, which are based on RGB cameras only. Such RGB only camera networks tend to fail in poor lighting conditions or dark environments. To the best of our knowledge, ours is the first attempt at cross-modal person re-identification where edge gradient features for local shape descriptors are used the same for both modalities. We have also exploited an effective metric learning approach to obtain a better re-identification matching score across the RGB and depth modalities. Experimental results on two benchmark RGB-D person re-identification datasets show the effectiveness of our proposed approach for the cross-modal re-identification problem.

Chapter 4

Depth Guided Attention for Person Re-identification in Multi-modal Scenario

4.1 Introduction

In recent years, Person re-identification (Re-ID) has gained great attention in both the computer vision community and industry because of its practical applications, such as in forensic search, multi-camera tracking and public security event detection. Person re-identification is still a challenging task in computer vision due to the variation of person pose, misalignment, different illumination conditions and diverse cluttered backgrounds. Fig. 4.1 shows a typical person re-identification system, where the task is to match the unknown probe with a set of known gallery images captured over non-overlapping cameras. It can be clearly observed from Fig. 4.1 that background clutter here in the scene works as the source of noisy information. And the trained model could be suffering from over-fitting as noisy information could propagate to it as salient features.

State-of-the-art approaches in Re-ID deal with this problem by relying on different attention-based mechanisms [63, 64, 65, 66]. All state-of-the-art attention-based Re-ID approaches can be placed into two categories: whole body attention and part-based attention. In the former case, methods focused on whole body attention, fully focused on the foreground while part-based methods focus more on local body parts. In all of these cases, methods rely on complex dedicated architectures which hinder the processes to deploy them in real world applications due to their large and over-parametrized models. Moreover, these methods are mainly based on RGB input that do not leverage additional information from other sources such as depth images.

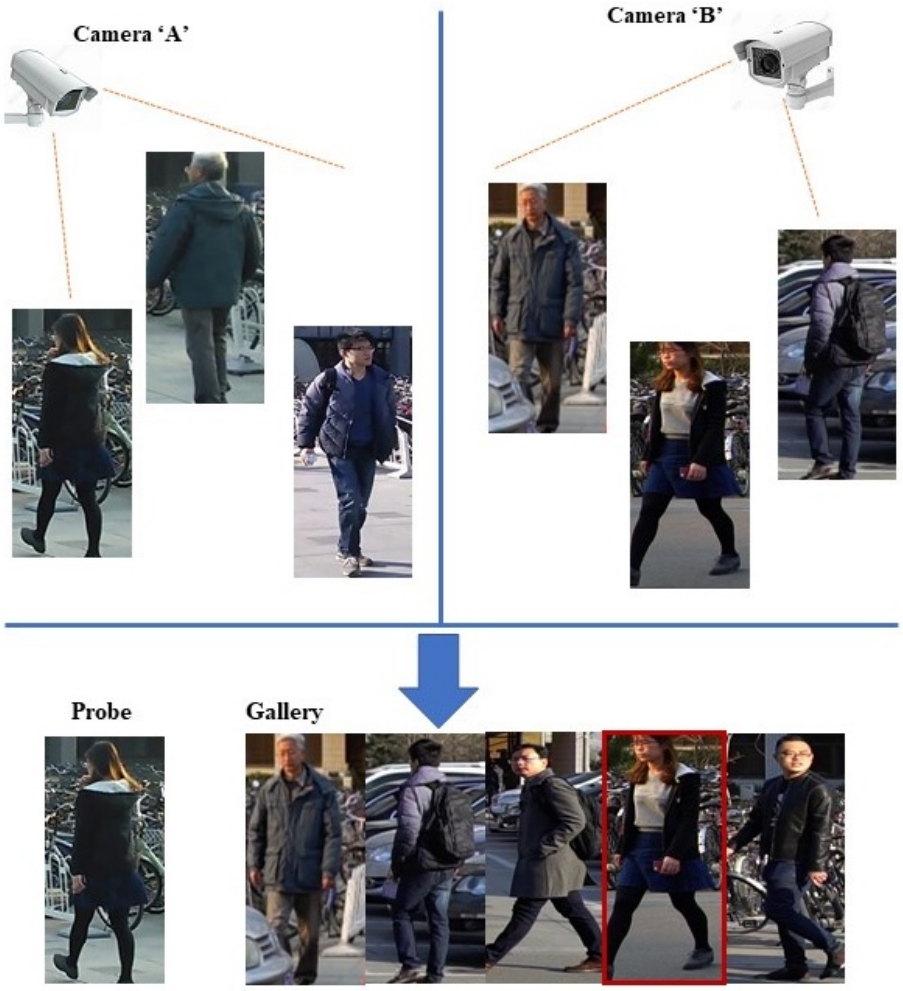

Figure 4.1: Illustration of challenges for a typical re-identification system. Sample images are taken from [80].

In our work, we emphasize how to extract discriminative and robust features using a depth sensor-based camera (e.g. Microsoft Kinect) when an individual appears on different cameras with diverse cluttered backgrounds. Specifically, whenever videos are recorded with a Kinect camera (i.e. RGB-D sensor) for each person, the Kinect SDK provides RGB frames, depth frames, the person's segmentation mask and skeleton data [49] with low computational effort.

In this research, we introduce depth guided binary segmentation masks to construct masked-RGB images (i.e. foreground images), where masked-RGB images retain the whole-body part of a person with different viewpoint variations and pose (see Fig. 4.2). In this work, we also focus on long-term person re-identification for RGB-D sensors with different pose variations of a person, which is suited to our

proposed approach.

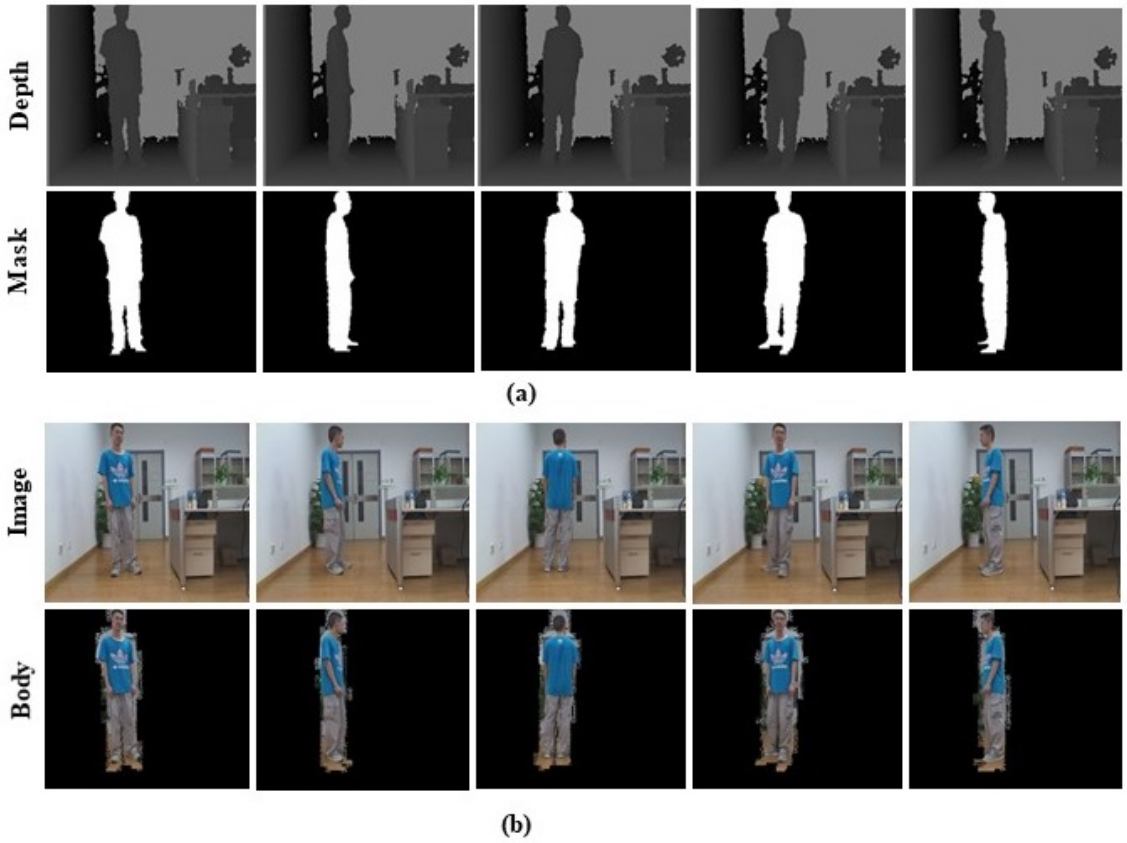

Figure 4.2: (a) Illustration of depths and their corresponding masks. (b) Examples of RGB images [58] and their corresponding body regions extracted directly with the masks.

Most previous methods directly learn features from the whole image which contain a person's body with a cluttered background. Recently, several deep learning methods have been proposed to learn features from the body parts [30] and pose[84, 85]. These methods have been proved effective through extracting features exactly from the body part rather than the background regions in the person image (i.e. pedestrian bounding box). It indicates that eliminating the background clutter in each person image is helpful for improving the performance of person re-identification.

This work also proposes a new deep learning Re-ID framework that takes into ac-count the additional information from the depth domain, thanks to the depth camera. Unlike past methods, the proposed approach exploits the advantage of using the depth image to generate a person's segmentation mask that helps us to develop deep learning methods which focus only on the foreground.

We evaluated the proposed method on the publicly available RGB-D dataset Robot-PKU RGBD-ID. Experimental results show the effectiveness of our proposed method. The contributions of this work can be summarized as follows:

1. We introduce a depth guided (DG) attention-based person re-identification framework. The key component of this framework is the depth-guided fore-ground extraction that helps the model to dynamically select the more relevant convolutional filters of the backbone CNN architecture.

2. Extensive experiments show the effectiveness of the proposed method in a depth-based benchmark re-identification dataset.

4.2 Methodology

In this section, we present our proposed depth guided attention-based person Re-ID in detail. First, we describe the overall framework of our method, then we present our triplet-based convolutional neural networks (CNNs) structure.

4.2.1 The Overall Framework

Our proposed pipeline is illustrated in Fig. 4.3. Our Re-ID framework consists of two states: depth guided body segmentation and triplet loss for re-identification.

In the first stage, we extract the foreground part of each image with the help of depth guided person segmentation masks. Once the foreground has been separated, then we feed the extracted body part T into the CNN model for feature mapping. For a given mask I_m and corresponding RGB image I_{rgb}, we separate the foreground after performing following operation,

$$T = I_m \otimes I_{rgb} \tag{4.1}$$

where $\otimes$ represents the element-wise product.

In the second stage, we describe the whole training procedure for Re-ID with CNN blocks. All the CNN blocks share parameters (i.e. weights and biases). During training, three CNNs take triplet examples (i.e. three foreground images), which is denoted as $T_i = (T_i^a, T_i^p, T_i^n)$ and forming the *i-th* triplet, where superscript 'a' indicates the anchor image, 'p' indicates positive image and 'n' indicates negative image. 'a' and 'p' come from the same person while 'n' is from a different person.

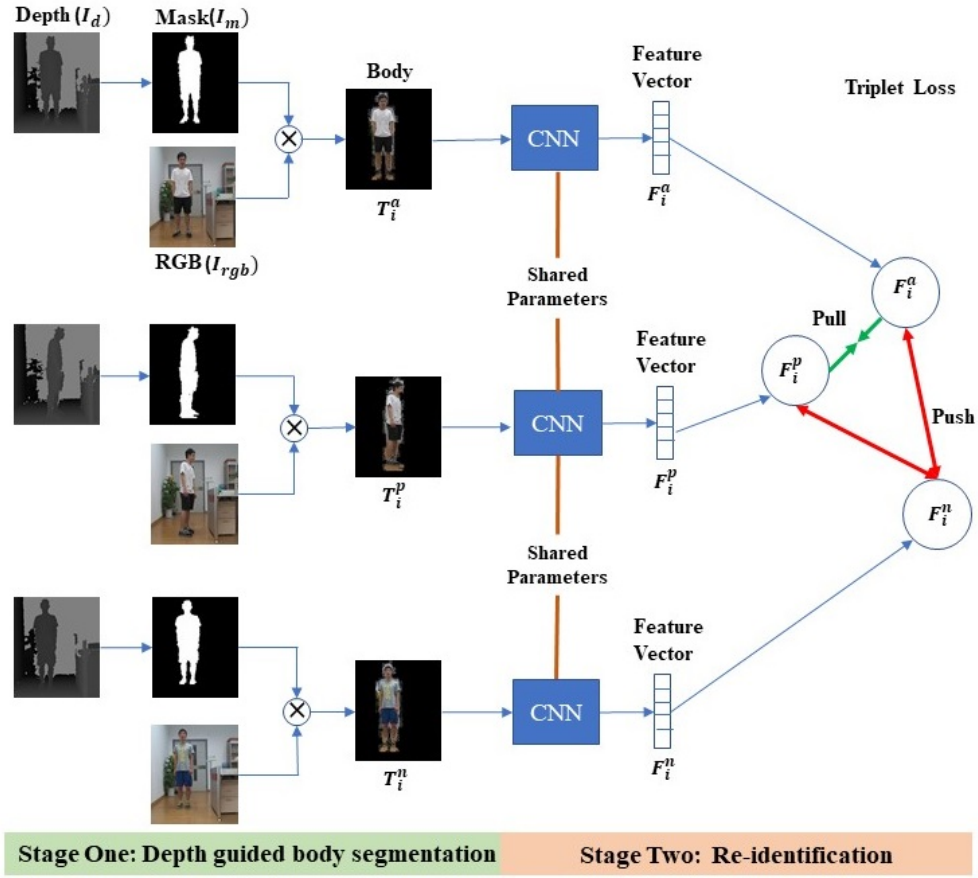

Figure 4.3: Triplet training framework for re-identification. It is composed of two stages: 1) Depth guided body segmentation and 2) Body segmented images are fed into three CNN models with shared parameters, where the triplet loss aims to pull the instances of the same person closer and at the same time, push the instances of different persons farther from each other in the learned feature space.

Foreground images are fed into the CNN model and maps the triplets T_i from the raw image space into a learned feature space $F_i = (F_i^a, F_i^p, F_i^n)$. For details, when a sample image is fed into the CNN model, it maps to the deep feature space $F = \varphi(x)$, where $\varphi(\cdot)$ represents the mapping function of the whole CNN model and x is the input representation of the corresponding image T.

4.2.2 Triplet Loss

The CNN model is trained with triplet loss function introduced by Weinberger and Saul [92]. In particular, the triplet loss has been shown to be effective in state-of-the-art person Re-ID systems [26, 81]. The triplet loss function aims to reduce the distance of feature vectors (i.e. F_i^a and F_i^p) taken from the same person (i.e. a and

p) and enlarge the distance between different persons (i.e. a and n). It is defined as

$$L_{trp} = \max\left\{0, \left\|F_i^a - F_i^p\right\|_2^2 - \left\|F_i^a - F_i^n\right\|_2^2 + m\right\} \tag{4.2}$$

where $\|\cdot\|_2^2$ is the squared Euclidean distance and m is a predefined margin which regularizes the distance. In our work, we train our model with margin $m = 0.3$. We use the Euclidean distance in our all experiments because the authors in [26] notice that using the squared Euclidean distance makes the optimization more prone to collapsing, whereas using an actual (non-squared) Euclidean distance is more stable.

Triplet generation is crucial to the final performance of the system. When the CNN is trained with the triplet inputs for a large-scale dataset then there can be an enormous possible number of combinations of triplet inputs (because triplet combinations increase cubically), making the training of all possible triplets impractical. To address this issue, we follow the Batch-hard triplet mining strategy introduced in [26]. The main idea is to form a batch by randomly sampling P identities and then randomly sampling K instances from each identity, and thus a resulting mini-batch contains $P \times K$ images in total. The Batch-hard triplet loss (BHtrp) can be formulated as

$$L_{BHtrp} = \sum_{i=1}^{P}\sum_{a=1}^{K}\left[m + \max_{p=1\ldots K}\left\|F_i^a - F_i^p\right\|_2 - \min_{\substack{n=1\ldots K \\ j=1\ldots P \\ j\neq i}}\left\|F_i^a - F_j^n\right\|_2\right]_+ \tag{4.3}$$

where F_i^a, F_i^p and F_i^n are normalized features of anchor, positive and negative samples respectively, and $\left[\cdot\right]_+ = \max(., 0)$.

4.3 Experiments

In this section, we evaluate the performance of our approach by performing experiments on the RobotPKU RGBD-ID [58] dataset.

4.3.1 Dataset

There are some publicly available RGB-D datasets [49, 47] which are very small in size, making it difficult to train a good model using our deep learning approach. Therefore, we consider the RobotPKU RGBD-ID dataset because this dataset consists of a decent amount of instances and a large number of frames per instance with different pose variations. This dataset was collected with Kinect sensors using the

Microsoft Kinect SDK. There are 180 video sequences of 90 people, and for each person still and walking sequences were collected in two separate indoor locations.

Data Pre-processing. Depth sensor-based cameras can capture depth images of a person within a particular range. In situations where depth sensors cannot capture depth frames properly, our system cannot extract the foreground part of the RGB image (see Fig. 4.4). Therefore, in our experiment, we consider only those RGB frames that have proper depth images of a person which can generate proper masks. After pre-processing, we obtain about 7,109 frames for training and 6,958 frames for testing, which come from 46 and 44 different identities respectively. We note that this is not a serious limitation as our system still covers a wide range of real world use cases.

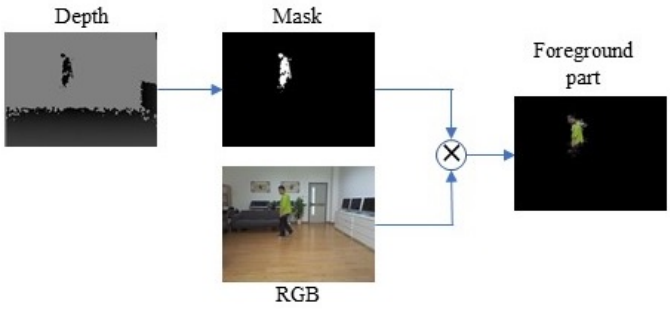

Figure 4.4: Illustration of the limitation of depth sensor to capture the depth frame of a distant person and their corresponding person segmentation mask.

4.3.2 Evaluation Protocol

We use cumulative matching characteristic (CMC) for quantitative evaluation, which is common practice in the Re-ID literature. For our experimental dataset, we randomly select about half of the people for training, and the remaining half for testing. In the testing phase, for each query image, we first compute the distance between the query image and all the gallery images using the Euclidean distance with the features extracted by the trained network, and then return the top n images which have the smallest distance to the query image in the gallery set. If the returned list contains an image featuring the same person as that in the query image at the *k-th* position, then this query is considered as rank k. In all our experiments, rank 1 result is reported.

4.3.3 Implementation Details

In our experiments, we use ResNet-18 [32] as well as ResNet50 [32] as the backbone CNN model. We use ResNet18 because it takes less memory and is computationally efficient, and the parameters are pre-trained on the ImageNet dataset [93]. Following the state-of-the-art methods, we also did our experiments using ResNet50. We train our model with stochastic gradient descent with a momentum of 0.9, weight decay of 5×10^{-4}, and initial learning rate of 0.01. The batch size is set to $32 \times 4 = 128$, with 32 different persons and 4 instances per person in each mini-batch. In our implementation, we follow the common practice of using random horizontal flips during training [21]. We resize all the images to 256×128. Our framework is implemented on the Pytorch [82] platform.

4.3.4 Experimental Evaluation

In this section, we report our experimental results on the RobotPKU RGBD-ID dataset. To demonstrate the effectiveness of our method using the additional information available from the depth domain, first we evaluate our proposed approach with different backbone architectures (such as ResNet50 and ResNet18) and variants of the original backbones. Second, we compare our approach with the available state-of-the-art methods for the given dataset.

Evaluation with different backbone. The goal of this experimental evaluation to check the effectiveness of our proposed method for different backbone architectures. As we already mentioned, we choose ResNet50 and ResNet18 as our backbone architectures. We also try different variants of those backbone architectures. To do so, we adopt the stride version of ResNet50 and ReNet18 by changing the stride of the last convolutional layer from 2 to 1, which basically increases the resolution of the final activation layers. We report our results in Table 4.1, and summarize the results using bar diagram in Fig. 4.5. Table 4.1 reports the rank-1 accuracy rate of the methods on the experimental dataset. We can make the following observations from these reported results:

ResNet50-strided indeed outperforms the original ResNet18 and ResNet18-strided for both scenarios in all the measures, which confirms our claims that increasing resolution on the final activation does affect the re-identification accuracy. The rank-1 performance improvement of the ResNet18-strided version over the original ResNet18 is 3.41% on both RGB and depth guided (DG) foreground images. From the above

Table 4.1: Comparison results of our method with different backbone architectures on RobotPKU dataset.

Method	**Backbone**	**Rank-1 (%)**
RGB	ResNet18	84.09
DG foreground	ResNet18	86.36
RGB	ResNet18-strided	87.50
DG foreground	ResNet18-strided	89.77
RGB	ResNet50-strided	90.90
DG foreground	ResNet50-strided	92.04

results, we can also see that our depth guided approach outperforms RGB for all the backbone CNN architectures.

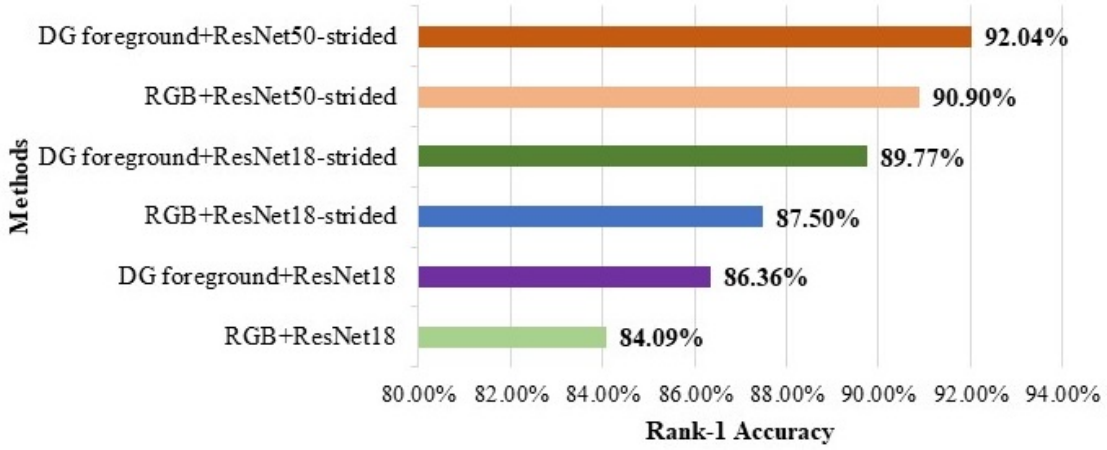

Figure 4.5: The performance of our method with different backbone networks on RobotPKU dataset.

Our proposed depth-guided foreground approach consistently works well for both versions of the considered backbone CNNs. The margin of improvement of our proposed approach considering original backbone architectures are relatively higher than their strided version. This implies that the finer details introduced by the proposed architecture on backbone architectures further improves the re-identification accuracy.

Comparison with Representative State-of-the-art Methods. The aim of these experiments is to analyze and compare the effectiveness of our proposed depth-guided foreground method to relevant state-of-the-art methods. Table 4.2 and Fig. 4.6 report the comparative performances of our methods with the state-of-the-art methods. Though some state-of-the-art methods [37, 83] performed experiments with this

dataset, but all of these are cross-modality matching (i.e. RGB-Depth matching). The performance of the cross-modality matching is very low, around 20%, that's why we do not include the results in this report.

In our experiments, we couldn't verify the results of the existing methods because public codes are not available, and it is also difficult and time consuming to implement the whole procedure.

Table 4.2: Comparison with other existing methods on RobotPKU dataset.

Method	**Rank-1** (%)
HSV [58]	69.79
SILTP [58]	46.71
Concatenation [58]	72.95
Score-level [58]	74.95
FFM [58]	77.94
RGB+ResNet18-strided **(Ours)**	87.50
DG foreground+ResNet18-strided **(Ours)**	89.77
RGB+ResNet50-strided **(Ours)**	90.90
DG foreground+ResNet50-strided **(Ours)**	92.04

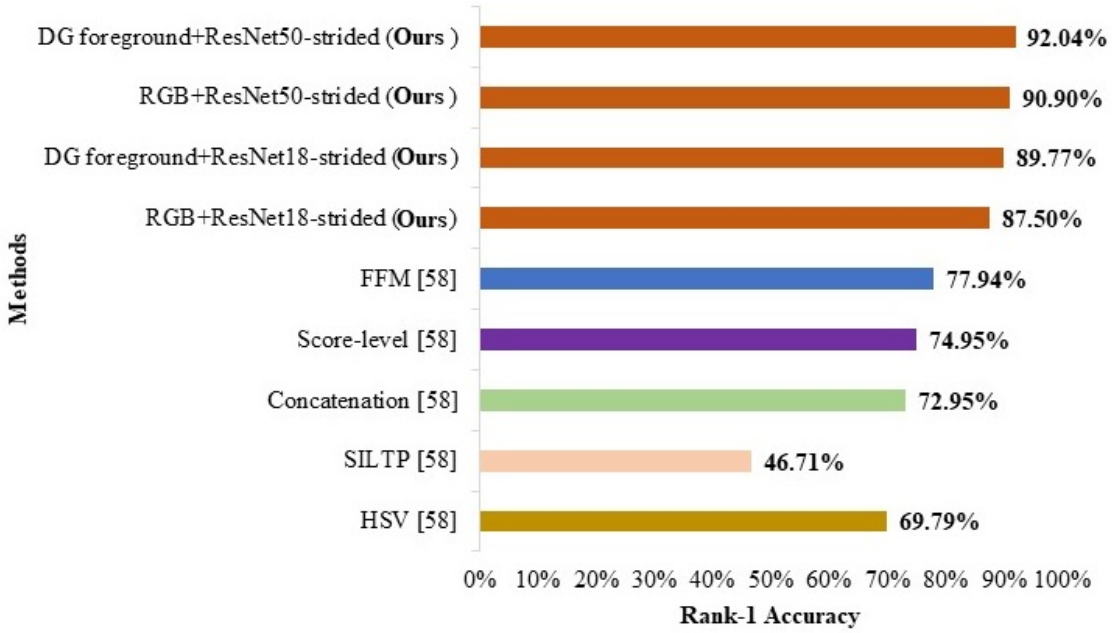

Figure 4.6: Comparison with different existing methods on RobotPKU dataset.

Our proposed approach considerably outperforms the state-of-the-art in all the measures. Among the alternatives, SILTP [58] performs worse while using hand-

crafted features which are mostly biased by the color or textures. The margin of improvement over the high performing state-of-the-art FFM (feature funnel model) is 14.1%. In FFM, the authors use both appearance and skeleton information provided by RGB-D sensors. The performance of the state-of-the-art methods varies significantly depending on their backbone architectures. We demonstrate the results of our method using different backbones and its variants in the previous section. Nevertheless, our proposed approach consistently outperforms the state-of-the-art methods irrespective to their backbone architectures.

Our proposed approach does not rely on complex dedicated architectures for extracting foreground as it does in most of the state-of-the-art works. Thus, our proposed approach is computationally efficient and provides better recognition accuracy using depth data, which can be useful to deploy in real-time applications.

4.4 Conclusion

In this work, we have presented a depth guided attention-based re-identification system. The key component of this framework is the depth-guided foreground extraction that helps the model to dynamically select the more relevant convolutional filters of the backbone CNN architecture, for enhanced feature representation and inference. Our proposed framework requires minimal modification to the backbone architecture to train the backbone network. Experimental results with a particular implementation of the framework (Resnet50 and Resnet18 with triplet loss) on the benchmark dataset indicate that the proposed framework can outperform related state-of-the-art methods. Moreover, our proposed architecture is general and can be applied with a multitude of different feature extractors and loss functions.

Chapter 5

Fusion in Dissimilarity Space for RGB-D Person Re-identification

5.1 Introduction

Person re-identification (Re-ID) is one of the most important parts of intelligent surveillance systems, which can recognize an individual across non-overlapping camera views. Person re-identification is a challenging task in computer vision because the visual appearance of an individual changes due to the variations in viewing angle, illumination intensity, pose, occlusion and background clutter. There have been a number of proposed approaches to address these problems based on conventional RGB cameras [86, 24, 87, 88, 12, 89, 23, 90] and recent invented modern RGB-D sensors [54, 46, 47, 48, 49, 50, 91]. Based on conventional RGB cameras, Re-ID researchers perform RGB-RGB matching, which is the most common scenario. While the RGB modality has been widely used, other modalities (i.e. depth and skeleton) can also be used as an additional information by taking advantage of RGB-D sensors to tackle some constraints (e.g. illumination), and form robust features by combining with visual features (i.e. RGB).

Most of the above mentioned RGB-D sensor-based Re-id works propose hand-crafted methods to extract new types of features from depth and skeleton joint points. These types of features are invariant against many variations such as illumination changes. Some Re-ID researchers combine these features with appearance features to enhance the Re-ID accuracy using feature-level fusion [55] and score-level fusion [56, 57, 58] techniques. In the most recent literature, some researchers have started to use deep learning methods for RGB-D person Re-ID [59, 60, 61]. These deep learning Re-ID approaches combine RGB-D sensor-based multi-modal features using the feature-level fusion strategy (see Fig. 1(a)) [59, 60] where [59] uses a multi-modal fusion layer

to fuse depth and RGB appearance features, and [60] designs a uniform and variational multi-modal auto-encoder at the top layer of their proposed deep network.

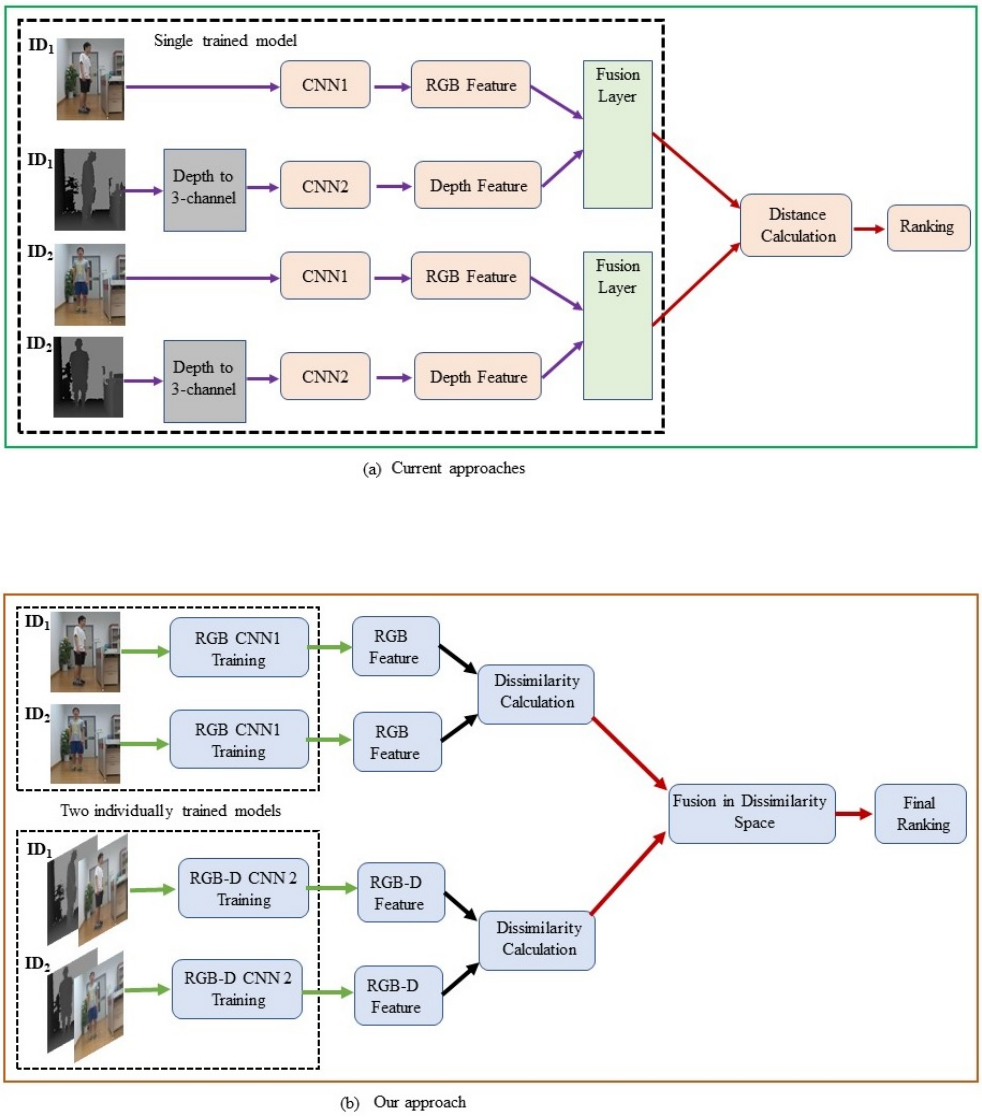

Figure 5.1: (a) A schematic of typical re-identification frameworks with deep learning. Current approaches focus on a feature-level fusion strategy with a single trained model. (b) Different from them, we use two individual trained models to extract features from 3-channel RGB and 4-channel RGB-D images accordingly.

These approaches, however, use a single trained model for multi-modal features (i.e. RGB and Depth) where they use 3-channel RGB and processed depth images (i.e. converted to 3-channel image) to increase Re-ID performance. In [61], two CNN streams (RGB CNN and depth CNN) separately process RGB image and depth image, and then features from the last fully connected layer of the both CNNs are

fused for jointly learning Re-id framework. Although these approaches achieve higher re-identification accuracies, the feature-level fusion may lead to the model being over-fitted as the fusion of noisy/heterogeneous features result in the noisy parts of the features to be dominated in the decision process. In our work, we address this issue by leveraging the fusion in dissimilarity space for multi-modal images (i.e. RGB-D) to increase the re-identification accuracy.

In this work, we focus on two individual modes instead of a single mode for RGB-D person re-identification (see Fig. 5.1). Unlike most existing learning-based RGB-D person re-identification methods which exploit the RGB and depth information from two different channels but fuse in a single fusion layer under a joint learning framework, we emphasize on two individually trained models based on 3-channel RGB and 4-channel RGB-D images (see Fig. 5.2), and compute the dissimilarity between query (i.e. RGB/RGB-D) and gallery (i.e. RGB/RGB-D) using feature embeddings extracted from two different trained models. The calculated dissimilarities for two individual modes are then fused in dissimilarity space to obtain final matching scores between query and gallery.

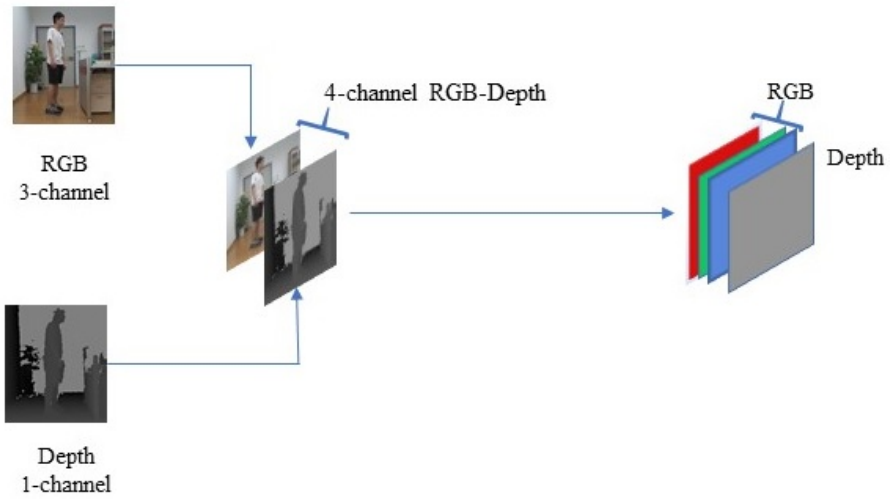

Figure 5.2: Formation of a 4-channel RGB-D image for person Re-ID input.

In this work, we take RGB and depth information in the form of RGB and RGB-D for two individual models. Therefore, we have the privilege to get the ensemble of RGB and RGB-D based trained models in dissimilarity space. Ensembling in such scenarios helps us to overcome the overfitting problem, while conventional feature

fusion approaches may suffer from overfitting due to the fusion of noisy/heterogeneous feature points.

Generally, depth information is robust to the variations of illumination, viewpoint, and resolution. In our work, we use RGB-D images which contain one more channel of depth information when compared with RGB images (see Fig. 5.2) and exploit the advantage of having an extra channel in the form of an illumination invariant depth image, and we also adjust the 4-channel RGB-D input with a 4-channel adaptive CNN in our Re-ID framework.

The main contributions of this work are as follows:

- First, we propose a novel Re-ID technique that exploits the advantage of using multi-modal data for fusing in dissimilarity space, where we design a 4-channel RGB-D image input in the Re-id framework.

- Second, we present an RGB-D Re-ID dataset including 58 identities. For each identity, a sequence of RGB and depth images is captured by the Intel RealSense Depth Camera D435 [9] in three different indoor locations with different illumination conditions.

- Finally, experimental analysis on our proposed dataset and two publicly available datasets indicate that fusion in dissimilarity space assists to increase the recognition accuracy compared to fusion in feature space

The remainder of the chapter is organized as follows. In section 5.2, we describe our dissimilarity-based Re-id framework using 3-channel RGB and 4-channel RGB-D sensor data, and our collected dataset SUCVL RGBD-ID is described in section 5.3. In section 5.4, Th experimental results of our method on different datasets are reported and compared with state-of-the-art methods. The general observations and typical failure cases are discussed in section 5.5. Finally, we offer concluding remarks in section 5.6.

5.2 Methodology

In this section, we present our proposed person re-identification method. Our proposed method is illustrated with a flowchart in Fig. 5.1(b). We divide our whole Re-ID framework into two phases. In the first phase, we train two models M_1 and M_2 using RGB and RGB-D images, respectively on the same training dataset. We refer

to these models as "RGB CNN" and "RGB-D CNN", respectively. Specifically, RGB-D CNN takes RGB images and their corresponding depth images to form 4-channel images as input. In the second phase, we calculate dissimilarity scores between the probe and galleries on the same testing dataset for each individually trained model and then finally fuse both scores in dissimilarity space.

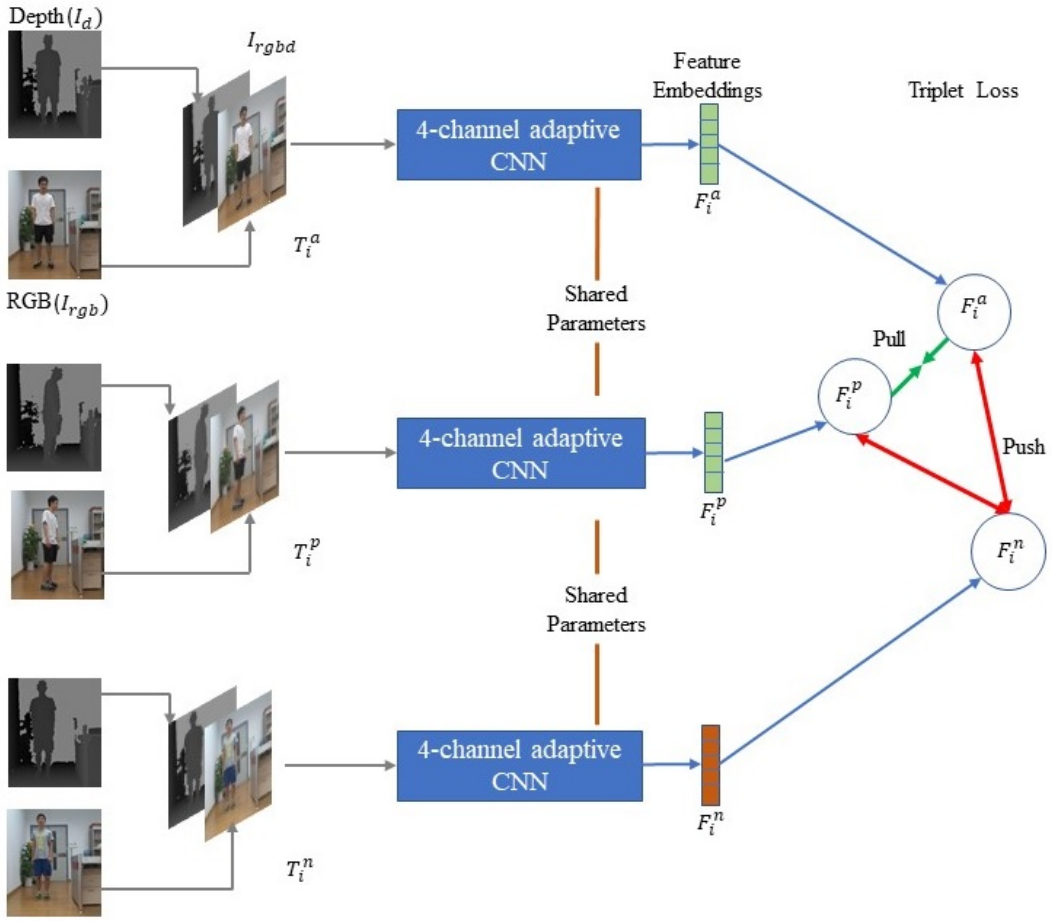

Figure 5.3: Triplet training framework of re-identification. It is composed of two stages: 1) 4-channel image formation with 3-channel RGB and 1-channel depth image and 2) 4-channel images are fed into three 4-channel adaptive CNN models with shared parameters, where the triplet loss aims to pull the instances of the same person closer and at the same time, push the instances of different persons farther from each other in the learned embedding space.

5.2.1 Model Training

In our proposed approach, RGB images are fed into three deep CNNs with shared parameters and triplet loss is used to train the RGB CNN. We use ResNet50 [32] as

the backbone for the RGB CNN and parameters are pre-trained on the ImageNet [93]. Ideally, ResNet50 accepts 3-channel input, but our Re-ID framework also needs to take 4-channel inputs. In Fig. 5.3, we present our Re-ID training framework with 4-channel RGB-D image input. 3-channel RGB images can be easily used with conventional pre-trained CNN models. But we require 4-channel RGB-D images as input to models with shared parameters, also pre-trained on the ImageNet. So, we modify the first convolution layer (by adding an extra 2D Conv layer) of ResNet50 in order to feed the model with 4-channel RGB-D images (see in Fig. 5.4).

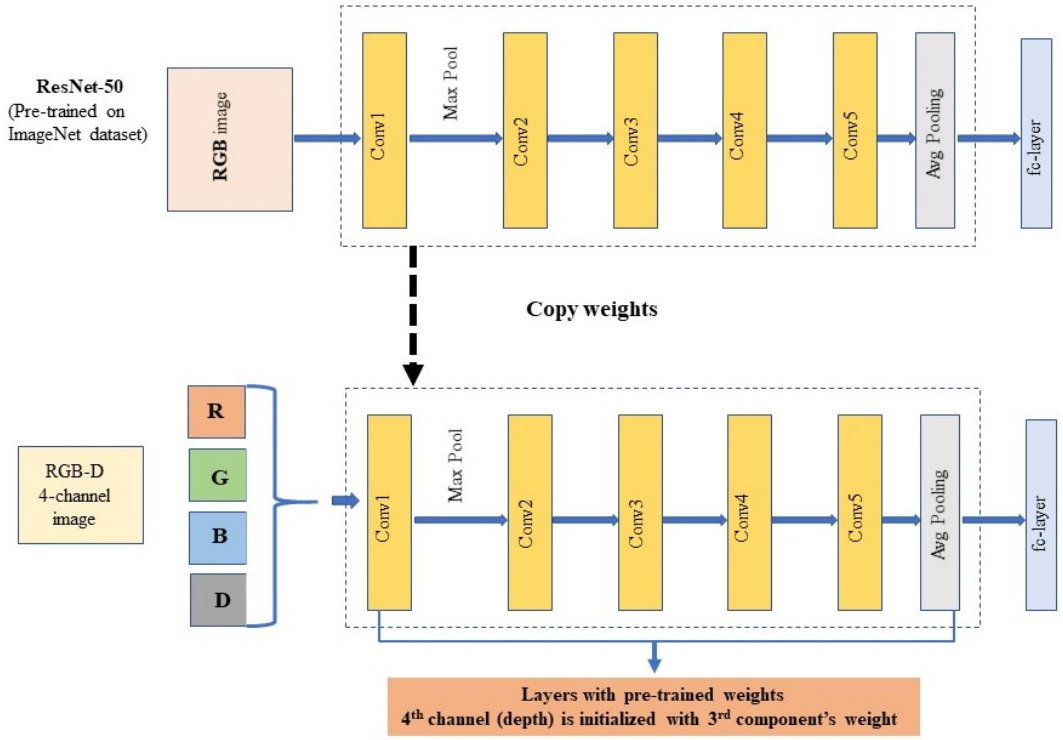

Figure 5.4: Adaptation of ResNet50 to 4-channel RGB-D image input.

Generally, ResNet50 should be first pre-trained on the ImageNet dataset to initialize the large numbers of parameters. In this work, we copy the parameters of the layers of the RGB model and then fine tune the RGB-D model with the same weights (w) of the RGB channels and 4th channel (depth channel) is initialized with the 3rd component's weights (see Fig. 5.4) to start the network training.

As like the RGB CNN model, we also train the RGB-D CNN model with triplet loss function. We describe the whole training procedure with three 4-channel adaptive CNN blocks (see Fig. 5.3) where all the CNN blocks share parameters (i.e. weights and biases). For a given RGB image I_{rgb} and corresponding depth image I_d, we

create a 4-channel RGB-D image I_{rgbd} as input. During training, three 4-channel adaptive CNNs take triplet examples (i.e. three I_{rgbd} images), which is denoted as $T_i = (T_i^a, T_i^p, T_i^n)$ and forming the *i-th* triplet, where superscript 'a' indicates the anchor image, 'p' indicates *hard positive* image and 'n' indicates *hard negative* image. 'a' and 'p' come from the same person while 'n' is from a different person. RGB-D images are fed into the 4-channel adaptive CNN model and maps the triplets T_i from the raw image space into a learned embedding space $F_i = (F_i^a, F_i^p, F_i^n)$. For details, when a sample image is fed into the CNN model, it maps to the feature embedding space $F = \varphi(x)$, where $\varphi(\cdot)$ represents the mapping function of the whole CNN model and x is the input representation of the corresponding image I_{rgbd}. For each image in the triplet examples, we calculate the gradient using the values of $\varphi(F_i^a)$, $\varphi(F_i^p)$, $\varphi(F_i^n)$ and $\frac{\delta\varphi(F_i^a)}{\delta w}$, $\frac{\delta\varphi(F_i^p)}{\delta w}$, $\frac{\delta\varphi(F_i^n)}{\delta w}$, which can be obtained by separately running the standard forward and backward propagation.

The RGB-D CNN as well as RGB CNN networks are trained using a triplet hard loss technique. In this technique, when a network is trained, the triplet loss function reduce the distance of feature embeddings (i.e. F_i^a and F_i^p) taken from the same person (i.e. anchor 'a' and *hard positive* 'p') and enlarges the distance between different persons (i.e. anchor 'a' and *hard negative* 'n') (see Fig. 5.3).

Triplet generation is an important factor to the final performance of the system. We follow Batch-hard triplet mining strategy, similar to our previous method, to tackle the generation of unnecessary number of triplet inputs. A batch is formed by randomly sampling P identities and then randomly sampling K instances from each identity, and thus a resulting mini-batch contains $P \times K$ images in total. The Batch-hard triplet loss (BHtrp) can be formulated as

$$L_{BHtrp} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\ anchors} \Big[m + \overbrace{\max_{p=1\ldots K} \left\|F_i^a - F_i^p\right\|_2}^{hardest\ positive} - \overbrace{\min_{\substack{n=1\ldots K \\ j=1\ldots P \\ j\neq i}} \left\|F_i^a - F_j^n\right\|_2}^{hardest\ negative}\Big]_+ \tag{5.1}$$

where F_i^a, F_i^p and F_i^n are normalized feature embeddings of anchor, positive and negative samples respectively, m is predefined margin and $\left[\cdot\right]_+ = \max(., 0)$.

Our whole training procedure is shown in **Algorithm 1**, which goes through all the triplets in each mini-batch to accumulate the gradients for each iteration and obtain model M_2 for RGB-D images.

Algorithm 1 shows the overall implementation of our training procedure for RGB-D images.

Input: Training samples of 4-channel RGB-D images $\{T_i\}$. Initialize the learning rate μ, margin m, network parameters $\{w\}$ and the number of iteration $t \leftarrow 0$.
Output: Model M_2

while $t < T$ **do**
  $t \leftarrow t + 1$
  $\frac{\delta L_{BHtrp}}{\delta w} = 0$
  Form all training triplet samples T_i from randomly sampled P identities and randomly sampled K instances from each identity;
  for all the training triplet samples T_i **do**
    calculate $\varphi(F_i^a)$, $\varphi(F_i^p)$, $\varphi(F_i^n)$ by forward propagation;
    calculate $\frac{\delta\varphi(F_i^a)}{\delta w}$, $\frac{\delta\varphi(F_i^p)}{\delta w}$, $\frac{\delta\varphi(F_i^n)}{\delta w}$ by back propagation;
  end for
  update the parameters $w^t = w^{t-1} - \mu_t \frac{\delta L_{BHtrp}}{\delta w}$
end while
return M_2

As the same procedure, except initializing network parameters, we follow the Algorithm 1 to obtain model M_1 for RGB image inputs.

5.2.2 Fusion Technique

We calculate dissimilarity scores (i.e. a score represents the Euclidean distance between two samples) using feature embeddings extracted from both trained models (M_1 and M_2) for a given set of gallery (G) and query (q) images (see Fig. 5.5). Then we sum both dissimilarity scores using the score-level fusion strategy (as the most of existing works for multi-modal cases follow this rule) in dissimilarity space with a fusion weight α. The fusion strategy is formulated as

$$D_{Fusion}(q, G) = \alpha D_{rgb}(q, G) + (1 - \alpha) D_{rgbd}(q, G) \tag{5.2}$$

where $D_{rgb}(q, G)$ and $D_{rgbd}(q, G)$ are the dissimilarity scores calculated using RGB and RGB-D feature embeddings respectively between each query sample (q) and gallery set (G), and $D_{Fusion}(q, G)$ is the final score between each query sample (q) and gallery set (G). **Algorithm 2** shows the fusion technique in dissimilarity space.

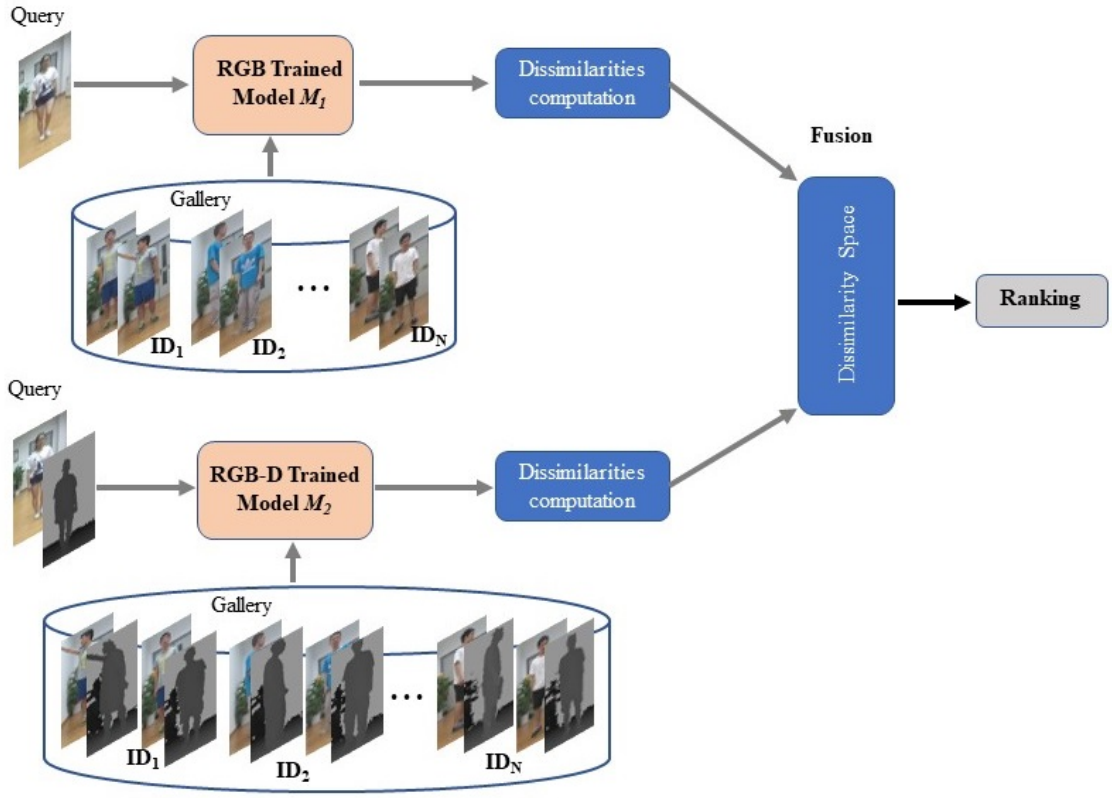

Figure 5.5: Final matching score calculation for our proposed Re-ID approach.

Algorithm 2

Input: Query q, gallery set G and initialize the dissimilarity fusion weight α.
Output: Dissimilarity scores $D_{Fusion}(q, G)$

1: Load model M_1, and extract query and gallery feature embeddings of RGB images;
2: **for** each RGB query image and gallery set **do**
3: calculate the dissimilarity scores using the following equation

$$D_{rgb} = \arg\min_{ID_i} D(q, ID_i), ID_i \in G$$

4: **end for**
5: Load model M_2, and extract query and gallery feature embeddings of RGB-D images;
6: **for** each RGB-D query image and gallery set **do**
7: calculate the dissimilarity scores using the following equation

$$D_{rgbd} = \arg\min_{ID_i} D(q, ID_i), ID_i \in G$$

8: **end for**

9: calculate final dissimilarity scores $D_{Fusion}(q, G)$ according to Eq. (5.2)

5.3 SUCVL RGBD-ID Dataset Description

In this section, we describe our collected RGB-D Re-ID dataset. To the best of our knowledge, there are five publicly available RGB-D datasets including RGBD-ID [50], KinectREID [55], BIWI RGBD-ID [49], IAS-Lab RGBD-ID [47] and RobotPKU [58] collected using the Microsoft Kinect Camera. All the above recorded datasets emphasize mainly on viewing angle variations. Some sequences were recorded in different lighting conditions in [47, 55]. Although most of these datasets are suitable for conventional RGB-D Re-ID methods, it is difficult to train a good model for deep learning methods because of the small size. Only RobotPKU dataset has a decent amount of instances and a large number of frames per instance with different viewpoint variations, though depth images are noisy (often body parts are absent in some frames). In our collected dataset, we emphasize on diverse lighting conditions in the recorded environments and have no alignment problem between RGB and depth images.

Our RGB-D Re-ID dataset contains 172 video sequences of 58 people col-lected using the Intel RealSense Depth Camera D435 and each person is captured under about 74 sequences of frames. Video sequences were recorded in three separate indoor locations on the same day, but different lighting conditions. The whole video recording scenario is depicted in the Fig. 5.6. Three cameras, indicated as Cam1, Cam2 and Cam3, were installed on the same floor of the building, but at three disjoint locations. To create lighting variations, Cam1 was installed in such a location where sunlight comes through two glass windows and changes the lighting condition of the environment. Cam2 was installed in our laboratory with an indoor lighting environment. The location of the third camera was in a corridor where indoor light was shut off, as a result the lighting condition was poor. All individuals were requested to walk normally forward to the camera. These videos were recorded at 30fps. The dataset includes synchronized RGB images (captured at a resolution of 1280×720 pixels) and depth images. Although the Intel RealSense Depth Camera D435 can capture images with a range up to 10m [9], we recorded all the videos within 5m ranges to get good quality depth images. The depth sensor can capture the depth information of each pixel by using infrared sensors, regardless of the pedestrian's color appearance and illumination condition in indoor environments.

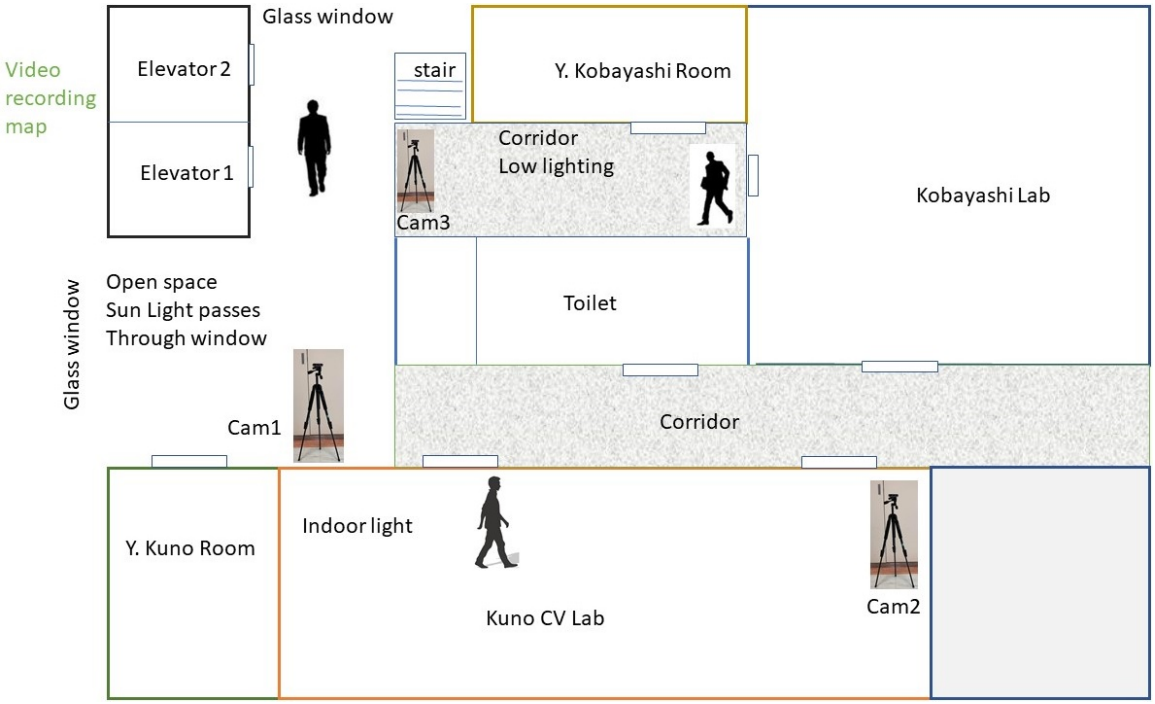

Figure 5.6: Overall video recording map.

As we setup up three cameras in three different illumination conditions, this makes it more challenging to recognize people from three non-overlapping cameras. We can see in Fig. 5.7, the RGB images of Cam1 are affected by sunlight which comes from outside through glass windows and changes the lighting environment of the indoor open space.

In Fig. 5.8, we show some example RGB and their corresponding depth images which were recorded in indoor and low lighting environments using Cam2 and Cam3 respectively. Although there are visual differences among RGB images for both cameras due to the lighting variations, depth images have no such differences (see in Fig. 5.7) because the depth sensors can capture illumination invariant high-quality depth images.

In our dataset, about half of the people wore jackets and some individuals wore face masks. Our dataset was designed for short-term person re-identification and therefore, the same person wore the same clothes in different acquisitions.

Figure 5.7: Example of RGB and their corresponding depth images. All images are captured on the same day and location but different times. Columns 1, 2, and 3 show the same person at different distances of view in normal lighting. Columns 4 and 5 show another person when sunlight comes through a glass window at a different time of the same day.

Figure 5.8: Columns 1, 2, and 3 show RGB and corresponding depth images captured by Cam2 in indoor lighting conditions, and columns 4 and 5 show the same person in low lighting environments captured by Cam3 at a different indoor location.

5.4 Experiments

Our proposed approach is evaluated on three RGB-D Re-ID datasets: RGBD-ID [50], RobotPKU RGBD-ID [58]] and our new proposed SUCVL RGBD-ID dataset. Although there are some other RGB-D Re-id datasets available, we chose the RobotPKU and RGBD-ID datasets for our experimental evaluation because of their large sizes.

5.4.1 Datasets

RobotPKU RGBD-ID. This dataset was collected with Kinect sensors using the Microsoft Kinect SDK. There are 180 video sequences of 90 people, and for each

person still and walking sequences were collected in two separate indoor locations. However, in some sequences, some depth frames are noisy and often body parts are absent in the images. This might happen because depth sensor-based cameras can capture depth images of a person within a particular range. In situations where depth sensors cannot capture depth frames properly, we discard all those frames using pre-processing techniques introduced in our previous method [91]. Therefore, in our experiment, we consider only those RGB frames that have proper depth images of an individual.

RGBD-ID. This dataset contains RGB and depth data for 79 individuals, and each individual has four acquisitions (walking1, walking2, collaborative and backwards), one rear view (backwards) and three frontal views (walking1, walking2 and collaborative). In each acquisition, four or five RGB and 3D frames (3D point clouds) are provided for each individual. Some individuals change their clothes in different acquisitions. As we perform our experiment with 3-channel RGB images and 4-channel RGB-D images, first we compute depth values from all 3D frames.

5.4.2 Evaluation Protocol

We use the cumulative matching characteristic (CMC) curve and mean average precision (mAP) for quantitative evaluation, which is common practice in the Re-ID literature. For all experimental datasets, we randomly select about half of the people for training, and the remaining half for testing. In the testing phase, for each query image(RGB/RGB-D), we first compute the dissimilarity (dissimilarities are a vector of Euclidean distance) between the query image and all the gallery images (RGB/RGB-D) using the feature embeddings extracted by the trained network (RGB/RGB-D model) (see Fig. 5.5), and then fuse both scores (RGB and RGB-D) in dissimilarity space. Finally, our Re-ID system returns the top n images which have the lowest dissimilarity to the query image in the gallery set. If the returned list contains an image featuring the same person as that in the query image at the *k-th* position, then this query is considered as *rank k*. We repeat the experiments 10 times, and the average accuracies of rank 1, 5 and 10 are reported along with mAP. All results reported in this paper are under a single query setting.

5.4.3 Implementation Details

We apply data augmentation techniques for both models (RGB and RGB-D) to increase the dataset's variability and improve network performance. All images are

resized to 256×192. In our implementation, we follow the common practice of using random horizontal flips during training [21]. ResNet50 with pre-trained parameters on ImageNet is adopted as the backbone network for the RGB model and we mentioned earlier in our proposed method section (5.2) how to train the RGB-D model for 4-channel image input. We train both our models with stochastic gradient descent with a momentum of 0.9, weight decay of 5×10^{-4}, and initial learning rate of 0.01. In our work, we set margin m=0.3 in Eq. 5.1 to train both models. We use the Euclidean distance instead of squared Euclidean distance in all our experiments because the authors in [26] notice that using the squared Euclidean distance makes the optimization more prone to collapsing, whereas using an actual (non-squared) Euclidean distance is more stable. The batch size is set to $20 \times 4 = 80$, with 20 different persons and 4 instances per person in each mini-batch. We set dissimilarity fusion weight $\alpha = 0.5$ in Eq. (5.2). This work is also implemented on the Pytorch platform.

5.4.4 Experimental Evaluation

In this section, we report our experimental results on our own SUCVL RGBD-ID dataset and the two other datasets mentioned above. To demonstrate the effectiveness of our method, first we compare the results of our dissimilarity based fusion model with two baseline models (RGB and RGB-D) as well as feature-level fusion of them. Second, we compare our Re-id approach with the available state-of-the-art methods for the given datasets.

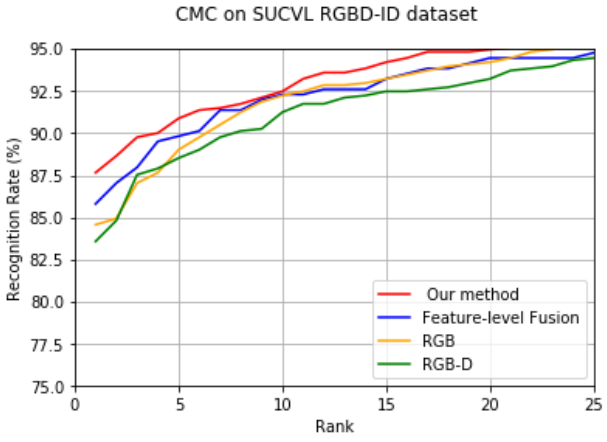

Figure 5.9: The CMC curve of different baseline methods and our approach on the SUCVL RGBD-ID dataset.

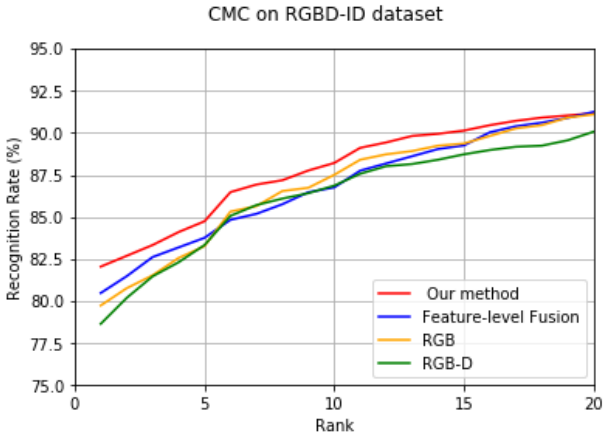

Figure 5.10: The CMC curve of different baseline methods and our approach on the RGBD-ID dataset.

Comparison with Baseline Models. The goal of this experiment is to check the effectiveness of our proposed method (Fusion in dissimilarity space) and compare with baseline models. The CMC curve of different baseline models and our method on the SUCVL RGBD-ID, RGBD-ID and RobotPKU datasets are shown in Fig. 5.9, 5.10 and 5.11, respectively. Table 5.1, 5.2 and 5.3 summarize the CMC curves reporting the rank-1, rank-5, rank-10 accuracies, and mAP for all the experimental datasets. From the CMC curves, we see that our proposed fusion model outperforms all the baseline models and feature-level fusion method especially at the top rank for all experimental datasets, which confirms our claims that multi-modal fusion in dissimilarity space increases the re-identification accuracy.

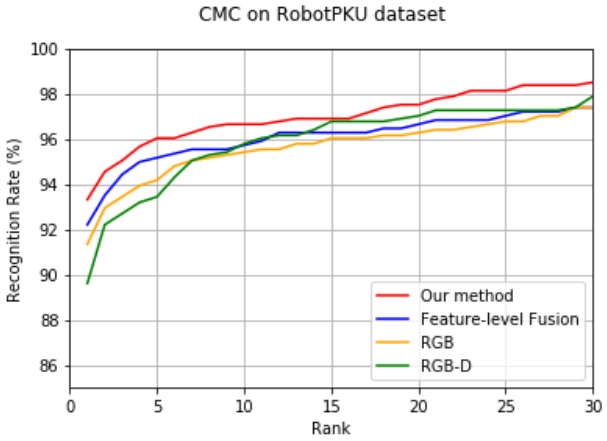

Figure 5.11: The CMC curve of different baseline methods and our approach on the RobotPKU dataset.

Table 5.1: Comparison results of our model with baseline models on SUCVL RGBD-ID dataset.

	SUCVL RGBD-ID			
Models	rank 1	rank 5	rank 10	mAP
RGB	84.56	89.01	92.22	71.14
RGB-D	83.58	88.51	91.23	70.11
Feature-level fusion	85.80	89.81	92.28	75.54
Ours fusion	**87.65**	**90.86**	**92.46**	**76.94**

Table 5.2: Comparison results of our model with baseline models on the complete RGBD-ID dataset.

	RGBD-ID			
Models	rank 1	rank 5	rank 10	mAP
RGB	79.74	83.27	87.50	69.11
RGB-D	78.65	83.33	86.86	68.32
Feature-level fusion	80.48	83.75	86.75	70.47
Ours fusion	**82.05**	**84.74**	**88.20**	**71.86**

Table 5.3: Comparison results of our model with baseline models on RobotPKU dataset.

	RobotPKU			
Models	rank 1	rank 5	rank 10	mAP
RGB	91.35	94.19	95.43	86.29
RGB-D	89.63	93.45	95.80	84.27
Feature-level fusion	92.22	95.18	95.73	87.56
Ours fusion	**93.33**	**96.04**	**96.66**	**89.49**

Table 5.1 shows that the mAP and rank-1 accuracy for the RGB model are 71.14% and 84.56%, and for the RGB-D model 70.11% and 83.58%, respectively. While our dissimilarity based fusion model increases the mAP to 76.94%, with 5.8% and 6.83% gain, and rank-1 accuracy to 87.65%, with 3.09% and 4.07% gain, respectively. Table 5.2 gives the comparison results on the complete RGBD-ID dataset where the RGB and RGB-D baseline models achieve 69.11% and 68.32% mAP, 79.74% and 78.65% rank-1 accuracy respectively. With the help of our fusion mechanism the mAP is increased to 71.86%, with 2.75% and 3.54% gain, and the rank-1 accuracy is increased to 82.05%, with 2.31% and 3.4% gain, respectively. Table 5.3 reports the results on

the RobotPKU dataset where mAP/rank-1 is 86.29%/91.35% for the RGB model and 84.27%/89.63% for the RGB-D model. Our fusion model improves the accuracy by +3.2%/+1.98% and +5.22%/3.7% for mAP/rank-1 over RGB and RGB-D baseline models accordingly. As shown in Tables 5.1, 5.2 and 5.3, the performance of our fusion approach is also better than the feature-level fusion method by considering the top rank and mAP for all experimental datasets.

Our proposed fusion approach consistently works well compared to the individual mode as well as feature-level fusion method for all experimental datasets. This implies that when the dissimilarity score vectors from two individual models (RGB and RGB-D) are fused in dissimilarity space, which increases re-identification accuracy.

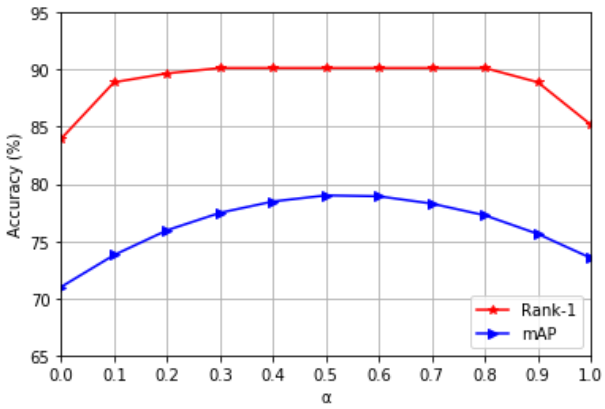

Figure 5.12: Effect of parameter α (shown by rank-1 and mAP accuracy) on the SUCVL RGBD-ID dataset.

Effect of Parameter α. For all experimental datasets, we repeated the experiments 10 times and estimated the average accuracies of rank 1, 5 and 10 along with mAP. To analyze the effects of dissimilarity fusion weight α in Eq. (5.2), we randomly choose one trial from the 10 trials and observe the effectiveness of α. we varied the value of α from 0 to 1 in the interval of 0.1 to see how the performance changed. The rank-1 accuracies and mAP under the different parameter settings are reported in Fig. 5.12, 5.13 and 5.14 for SUCVL RGBD-ID, RGBD-ID and RobotPKU datasets, respectively. It can be observed that, for all datasets, the rank-1 performance is improved significantly within range from $\alpha = 0.2$ to $\alpha = 0.6$, however SUCVL RGBD-ID extends the range to 0.8. Another evaluation measure, mAP, Fig. 5.12, 5.13 and 5.14 show that the best performance is approximately obtained when $\alpha = 0.5$ because mAP is calculated as the mean over all query images of the average precision. In our

experimental evaluation, we set $\alpha = 0.5$ to achieve the best performance of rank-1 and mAP.

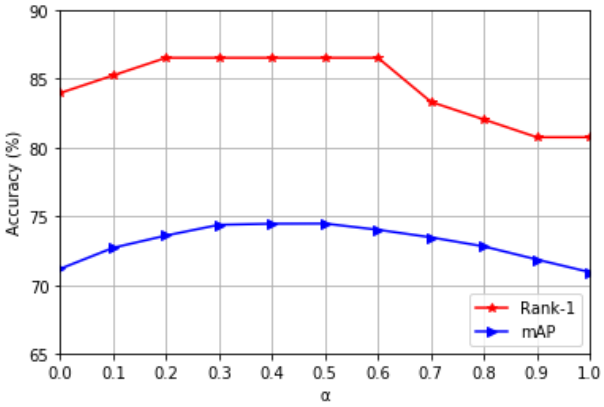

Figure 5.13: Effect of parameter α (shown by rank-1 and mAP accuracy) on the RGBD-ID dataset.

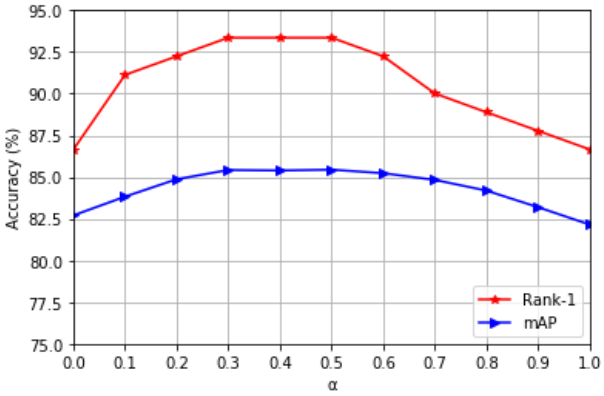

Figure 5.14: Effect of parameter α (shown by rank-1 and mAP accuracy) on the RobotPKU dataset.

Comparison with the State-of-the-art Methods. We further compare our proposed model with state-of-the-art methods on RGBD-ID and RobotPKU datasets.

RGBD-ID: On this dataset, we compare with several current proposed state-of-the-art methods. As this dataset has four different groups and few people wore different clothes in different acquisitions, some state-of-the-art methods disregarded those individuals who changed their clothes, and others considered the whole dataset in their experiments. In our experiment, we consider all the aspects for fair comparison with state-of-the-art methods.

Imani et al. [48, 94] divide the dataset into two smaller databases where the first one retains 59 people from walking1 and collaborative groups and the second one retains 72 people from walking2 and backwards groups after removing the people who changed their clothes in these groups. Table 5.4 reports the results with this setting and compares with other methods. Our method achieves high accuracy on rank-1 compared to other representative methods for both smaller datasets, although rank-5 and rank-10 fail to outperform. Table 5.4 shows the recognition rate of the walking1 and collaborative groups is slightly higher than the walking2 and backwards groups, this is because the first groups are in frontal view however, but latter groups are in the opposite view (frontal and back).

Table 5.4: Comparison of our proposed dissimilarity based fusion strategy with other methods when the RGBD-ID dataset is subdivided into two datasets by disregarding the people who changed their clothes. In all tables, * and '-' denote approximate values and non-present results respectively.

Methods	Walking1-collaborative				Walking2-backwards			
	rank 1	rank 5	rank 10	mAP	rank 1	rank 5	rank 10	mAP
SGLTrP3+score level [48]	76.58	-	99.35	-	72.58	-	95.91	-
GLVP3(Depth) + Skl. [94]	85*	**97***	98*	-	81*	**94***	**98***	-
RGB+depth [94]	81*	90*	**99***	-	72*	86*	92*	-
Ours Fusion	**87.58**	91.72	96.20	**88.8**	**84.95**	89.12	93.52	**85.8**

We also performed experiments on the whole RGB-D ID dataset. The performance of our proposed method and other state-of-the-art methods on this dataset is shown in Table 5.5. Our proposed fusion method in dissimilarity space for RGB-D Re-ID achieves 82.05% rank-1 accuracy and 71.86% mAP, outperforming the compared state-of-the-art methods. However, lower rank results of most of the state-of-the-art representative methods are better, but this aspect is negligible since top rank is the most significant for person re-identification tasks. The RTA [95], AIFL [96] and DVCov+SKL [46] methods consider the walking1 and walking2 groups for all individuals (79), where MMUDL [59] and UVDL [60] methods include the whole four groups (walking1, walking2, collaborative and backwards) for their experiments. As some individuals wear different clothes in different acquisitions in RGBD-ID dataset, some state-of-the-art methods ($MCMimpl^{DIS}$ multimodal [55] and APC-USG [54]) remove the corresponding tracks from different acquisitions and conducted experiments. To compare our proposed approach with the method presented in [54, 55], we

also remove the corresponding tracks and conduct experiments. Table 5.6 reports the results with this setting and compare with both methods. The rank-1 recognition rate of our method is 88.5%, against to the 77.7% of $MCMimpl^{DIS}$ multimodal [55]. An improvement of about 10.8% is achieved. Though, our approach fails to outperform APC-USG [54], but very close to the performance, only 0.84% lower.

Table 5.5: Comparison of our proposed dissimilarity based fusion strategy with other state-of-the-art methods when the entire RGBD-ID dataset is considered for experimental evaluation.

	RGBD-ID			
Methods	rank 1	rank 5	rank 10	mAP
RTA [95]	52.4	-	-	-
AIFL [96]	59.4	61	-	64.5
DVCov+SKL [46]	71.74	88.4	-	-
MMUDL [59]	76.7	87.5	96.1	-
UVDL [60]	76.7	**92**	**98.2**	-
Ours fusion	**82.05**	84.74	88.20	**71.86**

Table 5.6: Comparison of our proposed method with other state-of-the-art methods, where the people who change their clothes in different acquisitions, have been discarded from the computation.

	RGBD-ID			
Methods	rank 1	rank 5	rank 10	mAP
$MCMimpl^{DIS}$ multimodal [55]	77.7*	**94***	**99***	-
APC-USG [54]	**89.34**	-	-	-
Ours fusion	88.50	90.87	93.12	**84.09**

Table 5.7: Comparison with other methods on RobotPKU dataset.

	RobotPKU			
Methods	rank 1	rank 5	rank 10	mAP
HSV [58]	69.79	-	-	-
SILTP [58]	46.71	-	-	-
Concatenation [58]	72.95	-	-	-
Score-level [58]	74.95	-	-	-
FFM [58]	77.94	-	-	-
Depth Guided (DG) attention [91]	92.04	-	-	-
Ours fusion	**93.33**	**96.04**	**96.66**	**89.49**

As observable, our proposed score-level fusion for 3-channel RGB and 4-channel RGB-D in dissimilarity space, outperforms feature-level fusion methods like $MCMimpl^{DIS}$

multimodal [55], MMUDL [59] and UVDL [60] in the top rank, which indicates that fusion in dissimilarity space assists with increasing the recognition accuracy compared to the fusion in feature space.

RobotPKU: On RobotPKU dataset, our proposed method outperforms all the other methods and exceeds our previous model Depth Guided (DG) attention [91] by +1.29% in rank-1, results are reported in Table 5.7.

5.4.5 Runtime Performance Evaluation

We evaluate the running time of our re-identification system. We use the NVIDIA GeForce GTX TITAN X (GPU) to extract image features, and take SUCVL RGBD-ID dataset as an example. In the testing phase, we use 5,955 gallery and 81 query images which are taken from 27 individuals. As our method uses 3-channel RGB and 4-channel RGB-D images, we extract features separately using trained models M_1 and M_2. When we use 3-channel RGB images, it costs 47.573s to extract all features and 65.018s for 4-channel RGB-D images, in total 112.591s. Hence, our method takes 18.65×10^{-3}s in average to extract feature for each image. In addition, 81 queries are evaluated against all gallery images (5,955), it costs 0.2247s. So, our proposed method takes 2.77×10^{-3}s in average to obtain a rank list for each query image.

5.5 Discussion

In this section, we discuss the insights we observed in our experiments and typical failure cases.

5.5.1 General Observations

We make the following general observations from experimental results on our proposed dataset and two publicly available datasets.

- Among the different baseline models, feature-level fusion model achieves good accuracy even though this model may cause overfitting problem because of the direct fusion of two CNNs features extracted from two different modalities RGB and depth. However, blindly fusing of heterogeneous features may not increase discrimination power, as different features have different reliability. In contrast to fusion in feature space, ensembling of 3-channel RGB and 4-channel RGB-D based models in dissimilarity space mitigates the overfitting problem and

achieves superior accuracy which is shown in Fig. 5.9, 5.10 and 5.11. From these figures, we observe the effectiveness of our method for all experimental datasets. Beside two public datasets, we also verified our method using our own collected dataset (SUCVL RGBD-ID), which also outperforms the feature-level fusion method (see in Fig. 5.9), margin of improvement 1.85% and 1.4% for rank-1 and mAP, respectively.

- To verify the effectiveness of our method on different backbones of ResNet, we replaced the Res-Net50 with a shallower architecture (ResNet18) as a backbone network and performed the experiments on all experimental datasets. In this case, we kept the value of all parameters (margin, learning rate, momentum, weight decay and fusion weight) same as we set for ResNet50 and performed the experiments for given datasets. We observed the slight improvement of our fusion method over feature-level fusion for RobotPKU dataset, only 0.55% and 0.91% for rank-1 and mAP, respectively. For RGBD-ID dataset, results are almost same for our method and feature-level fusion for rank-1 and mAP. While experimented with SUCVL RGBD-ID dataset, we achieved slightly higher accuracy than feature-level fusion, 1.48% and 1.41% improvement for rank-1 and mAP, respectively. For the first two datasets, improvement was very low, however, for third one, improvement was slightly higher. This is because first two datasets were captured by Kinect depth sensor, which has limitation of capturing depth images within a particular range [97], while third one were captured by Intel RealSense depth camera which has wider range of capturing ability of good quality depth images [9]. Therefore, the general observation is that our dissimilarity based fusion method works well for deeper networks like ResNet50 than shallower network (ResNet18) for all experimental datasets.

5.5.2 Failure Cases Analysis

In this section, we illustrate the failure cases of our method, caused by two reasons: limitation of capturing depth image and dissimilarity fusion weight.

Our proposed method is two steps process. In the first step, two models are trained using 3-channel RGB and 4-channel RGB-D images. Depth sensor-based cameras can capture depth images of a person within a particular range, for example, operational range of Microsoft Kinect is between 0.8m and 4m. Although the Intel RealSense Depth sensor can capture image with a range up to 10 meters, but good quality depth image can capture within 6m. In situations where depth sensors cannot

capture depth frames properly (see Fig. 5.15), our proposed system fail to train a good model. Consequently, recognition rate drops significantly because feature embeddings extracted from RGB-D trained model for a given set of gallery and query images are not well formed.

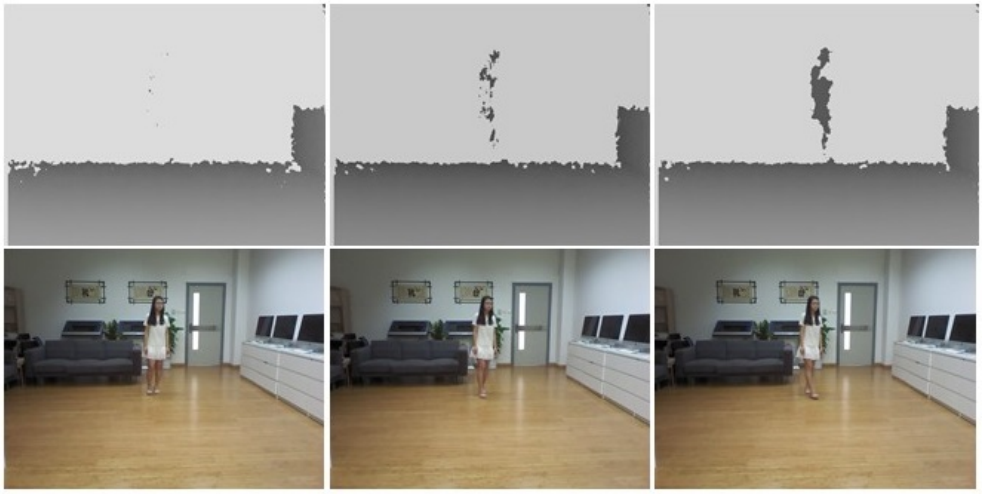

Figure 5.15: Illustration of failure case caused by depth sensor.

In the second case, if we set fusion weight $\alpha \leq 0$ or $\alpha \geq 1$ in Eq. (5.2), our method will fail.

5.6 Conclusions

In this work, we have presented a re-identification approach that exploits the advantages of having multi-modal images in the form of RGB-D. In this context, we developed an effective fusion technique in dissimilarity space for 3-channel RGB and 4-channel RGB-D images to increase re-identification accuracy. Most existing Re-ID approaches follow the feature-level fusion strategy, which may lead to the model being overfitted in the fusion of noisy/heterogeneous features, so there are chances of deterioration in the final recognition process. We have also proposed an RGB-D Re-ID dataset which was captured under diverse lighting conditions which makes it more challenging to recognize people. Experimental results on our collected dataset and two other benchmark datasets show the efficiency of our proposed approach for RGB-D person re-identification. Moreover, our proposed method is general and can be applied to a multitude of different RGB-D based applications.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we have presented three methods to tackle the challenges (i.e. very poor lighting conditions, clothing changes, background cluttered problem and also overfitting problem for multi-modal data.) of person re-identification with RGB-D sensors for video-surveillance. In the first method, we have investigated a cross-modal person re-identification to address the poor illumination and clothing changes problem with RGB-D sensors by making use of their depth data. In the second method, we have introduced a depth guided attention-based person re-identification approach to tackle the diverse clutter background problem in multi-modal scenario, and our third method is improved version of the previous method, where we fully use RGB and depth modalities to improve the re-identification accuracy.

In cross-modal person re-identification, we proposed a heterogeneous camera network where RGB cameras can capture RGB sequence of video frames in normal lighting conditions and depth cameras can capture illumination and color invariant depth sequence of video frames in very poor lighting conditions. To re-identify a person across RGB and depth modalities, we also proposed a body partitioning method and HOG based feature extraction technique on RGB and depth modalities. In addition, we have exploited a PCA and LDA based metric learning approach for person re-identification which has the ability to maximize inter-class variations and minimize the intra-class variations. A rigorous experimental analysis on two publicly available datasets, we have demonstrated the effectiveness of our cross-modal person re-identification method.

In multi-modal scenario, first we proposed a deep learning person re-identification framework in the form of attention mechanism to address diverse clutter background problem where we separated foreground part of an RGB image with the help of

depth-based additional information, unlike state-of-the-art methods of complex architectures. Second, we propose a novel Re-ID technique that exploits the advantage of using multi-modal data for fusing in dissimilarity space. In this method, we successfully adapt 4-channel RGB-D image inputs in the Re-ID framework.This adaption helps to use 4-channel RGB-D images and 3-channel RGB images separately to train two models, and then calculate the scores from each individual model. Finally, fuse the scores in the dissimilarity space which assists to overcome the overfitting problem due to the heterogeneous data as a result system performance increase.

6.2 Future Work

In this thesis, we have presented three methods to address the constraints, such as extreme low lighting conditions, clothing changes, background clutter, and also overfitting problem for multi-modal data. There are still many open issues associated with RGB-D and Infrared (IR) sensor-based person Re-ID problem, which we couldn't explore completely during this thesis period. Here, we mention two issues of multi-modal and cross-modal person Re-ID problem, which we want to extend.

- As RGB-D sensors provide RGB, depth and skeleton information simultaneously, therefore, we have a plan to extend our multi-modal fusion work for combining 4-channel RGB-D image features and skeleton-based features together to form a complete representation of human body.

- As new generation surveillance camera can automatically switch to infrared mode to capture the person image at night, therefore, we have a plan to extend our cross-modality person Re-ID problem for visible thermal, which will play important role in practical night-time video surveillance applications.

Publication List

[1] **M.K. Uddin**, A. Lam, H. Fukuda, Y. Kobayashi and Y. Kuno. Exploiting Local Shape Information for Cross-Modal Person Re-identification. In *Intelligent Computing Methodologies. Lecture Notes in Computer Science,* vol 11645, pages 74-85, Springer, 2019.

[2] **M.K. Uddin**, A. Lam, H. Fukuda, Y. Kobayashi and Y. Kuno. Depth Guided Attention for Person Re-identification. In *Intelligent Computing Methodologies. Lecture Notes in Computer Science,* vol 12465, pages 110-120, Springer, 2020.

[3] **M.K. Uddin**, A. Lam, H. Fukuda, Y. Kobayashi and Y. Kuno. Fusion in Dissimilarity Space for RGB-D Person Re-identification. *Array,* 12(2021)100089.

Bibliography

[1] Cisco visual networking index: Forecast and methodology, 2015/2020. *Cisco Systems Inc.*, 2016.

[2] CCTV. URL: https://www.pinterest.com/pin/653936808364956099

[3] J. Si, H. Zhang, C. Li and J. Guo. Spatial Pyramid-Based Statistical Features for Person Re-Identification: A Comprehensive Evaluation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* 48(7):1140-1154, 2018.

[4] L. Wei, S. Zhang, W. Gao and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 79-88, 2018.

[5] S. Ding, L. Lin, G. Wang and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition,* 48(10):2993-3003, 2015.

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

[7] H. Liu, L. Hu, and L. Ma. Online RGB-D person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology,* 2(1):48-55, 2017.

[8] S. Coşar and N. Bellotto. Human Re-Identification with a Robot Thermal Camera Using Entropy-Based Sampling. *Journal of Intelligent and Robotic Systems,* 98(1):85-102, 2020.

[9] Intel RealSense Technology Homepage, https://www.intelrealsense.com/depth-camera-d435/

[10] B. Ma, Y. Su and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing,* 32(6-7):379-390, 2014.

[11] A. Bhuiyan, A. Perina and V. Murino. Person re-identification by discriminatively selecting parts and features. In *European Conference on Computer Vision (ECCV),* pages 147-161. Springer, 2014.

[12] S. Liao, Y. Hu, X. Zhu and S.Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 2197-2206, 2015.

[13] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi and S.Z. Li. Salient color names for person re-identification. In *European conference on computer vision (ECCV),* pages 536-551. Springer, 2014.

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

[15] M.O. Almasawa, L.A. Elrefaei and K. Moria. A survey on deep learning-based person re-identification systems. *IEEE Access,* vol.7, pages 175228-175247, 2019.

[16] S. Liao and S.Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE international conference on computer vision (CVPR),* pages 3685-3693, 2015.

[17] X. Wang, W.S. Zheng, X. Li and J. Zhang. Cross-scenario transfer person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology,* 26(8):1447-1460, 2016.

[18] Y.C. Chen, W.S. Zheng, J.H.Lai and P.C. Yuen. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE transactions on circuits and systems for video technology,* 27(8):1661-1675, 2017.

[19] W.S. Zheng, S. Gong and T. Xiang. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence,* 35(3):653-668, 2013.

[20] D. Yi, Z. Lei, S. Liao and S.Z. Li. Deep metric learning for person re-identification. In *22nd International Conference on Pattern Recognition*, pages 34-39, 2014.

[21] E. Ahmed, M. Jones and T.K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 3908-3916, 2015.

[22] D. Cheng, Y. Gong, S. Zhou, J. Wang and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 1335-1344, 2016.

[23] R.R. Varior, M. Haloi and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision (ECCV),* pages 791-808, Springer, 2016.

[24] T. Xiao, H. Li, W. Ouyang and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 1249-1258, 2016.

[25] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng and S.Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *European conference on computer vision (ECCV),* pages 732-748, Springer, 2016.

[26] A. Hermans, L. Beyer and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737,* 2017.

[27] W. Chen, X. Chen, J. Zhang and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 403-412, 2017.

[28] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 6036-6046, 2018.

[29] F. Wang, W. Zuo, L. Lin, D. Zhang and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 1288-1296, 2016.

[30] D. Li, X. Chen, Z. Zhang and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 384-393, 2017.

[31] Z. Zheng, L. Zheng and Y. Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),* 14(1):1-20, 2017.

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

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 1-9, 2015.

[34] M. Geng, Y. Wang, T. Xiang and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244,* 2016.

[35] Y.J. Li, F.E. Yang, Y.C. Liu, Y.Y. Yeh, X. Du and Y.C. Frank Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 172-178, 2018.

[36] J. Zhuo, J. Zhu, J. Lai and X. Xie. Person re-identification on heterogeneous camera network. In *CCF Chinese Conference on Computer Vision,* pages 280-291, Springer, Singapore, 2017.

[37] F.M. Hafner, A. Bhuiyan, J.F. Kooij and E. Granger. RGB-depth cross-modal person re-identification. In *16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),* pages 1-8, 2019.

[38] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu and N. Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 13379-13389, 2020.

[39] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI),* pages 677–683, 2018.

[40] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR),* pages 5380–5389, 2017.

[41] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,* 2018.

[42] Mang Ye, Zheng Wang, Xiangyuan Lan and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI),* pages 1092–1099, 2018.

[43] G.A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang and Z.G. Hou. Cross-modality paired-images generation for RGB-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 34(07):12144-12151, 2020.

[44] G. Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR),* pages 3623–3632, 2019.

[45] Z. Wang, Zheng Wang, Y. Zheng, Y. Chuan and S. Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 618–626, 2019.

[46] A. Wu, W.S. Zheng and J.H. Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing,* 26(6):2588-2603, 2017.

[47] M. Munaro, A. Basso, A. Fossati, L. Van Gool and E. Menegatti. 3D reconstruction of freely moving persons for re-identification with a depth sensor. In *IEEE international conference on robotics and automation,* pages 4512-4519, 2014.

[48] Z. Imani and H. Soltanizadeh. Person reidentification using local pattern descriptors and anthropometric measures from videos of kinect sensor. *IEEE Sensors Journal,* 16(16):6227-6238, 2016.

[49] M. Munaro, A. Fossati, A. Basso, E. Menegatti and L. Van Gool. One-shot person re-identification with a consumer depth camera. In *S. Gong, M. Cristani, S. Yan, C. Loy. (eds) Person Re-Identification. Advances in Computer Vision and Pattern Recognition,* pages 161-181, Springer, London, 2014.

[50] I.B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani and V. Murino. Re-identification with rgb-d sensors. In *European conference on computer vision (ECCV),* pages 433-442, Springer, Berlin, Heidelberg, 2012.

[51] A. Haque, A. Alahi and L. Fei-Fei. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 1229-1238, 2016.

[52] Z. Imani, H. Soltanizadeh and A.A. Orouji. Tensor-based sparse canonical correlation analysis via low rank matrix approximation for RGB-D long-term person re-identification. *Multimedia Tools and Applications,* 79:11787–11811, 2020.

[53] Z. Imani and H. Soltanizadeh. Histogram of the node strength and histogram of the edge weight: two new features for RGB-D person re-identification. *Science China Information Sciences,* 61(9):092108, 2018.

[54] C. Patruno, R. Marani, G. Cicirelli, E. Stella and T. D'Orazio. People re-identification using skeleton standard posture and color descriptors from RGB-D data. *Pattern Recognition,* 89:77-90, 2019.

[55] F. Pala, R. Satta, G. Fumera and F. Roli. Multi-modal person re-identification using RGB-D cameras. *IEEE Transactions on Circuits and Systems for Video Technology,* 26(4):788-799, 2015.

[56] A. Mogelmose, C. Bahnsen, T. Moeslund, A. Clapés and S. Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 301-307, 2013.

[57] R. Kawai, Y. Makihara, C. Hua, H. Iwama and Y. Yagi. Person re-identification using view-dependent score-level fusion of gait and color features. In *Proceedings*

of the 21st International Conference on Pattern Recognition (ICPR), pages 2694-2697, Tsukuba, Japan, 2012.

[58] H. Liu, L. Hu and L. Ma. Online RGB-D person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology,* 2(1):48-55, 2017.

[59] L. Ren, J. Lu, J. Feng and J. Zhou. Multi-modal uniform deep learning for RGB-D person re-identification. *Pattern Recognition,* 72:446-457, 2017.

[60] L. Ren, J. Lu, J. Feng and J. Zhou. Uniform and variational deep learning for RGB-D object recognition and person re-identification. *IEEE Transactions on Image Processing,* 28(10):4970-4983, 2019.

[61] A.R. Lejbolle, K. Nasrollahi, B. Krogh and T.B. Moeslund. Multi-modal neural network for overhead person re-identification. In *International Conference of the Biometrics Special Interest Group (BIOSIG),* pages 1-5, Darmstadt, 2017.

[62] A.R. Lejbolle, B. Krogh, K. Nasrollahi and T.B. Moeslund. Attention in multi-modal neural networks for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 179-187, 2018.

[63] C. Song, Y. Huang, W. Ouyang and L. Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 1179-1188, 2018.

[64] D. Chen, S. Zhang, W. Ouyang, J. Yang and Y. Tai. Person search via a mask-guided two-stream cnn model. In *Proceedings of the European Conference on Computer Vision (ECCV),* pages 734-750, 2018.

[65] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak and M. Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 1062-1071, 2018.

[66] H. Cai, Z. Wang and J. Cheng. Multi-scale body-part mask guided attention for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2019.

[67] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 3431-3440, 2015.

[68] K. He, G. Gkioxari, P. Dollár and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision (CVPR),* pages 2961-2969, 2017.

[69] X. Liang, K. Gong, X. Shen and L. Lin. Look into person: Joint body parsing and pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence,* 41(4):871-885, 2018.

[70] R. Alp Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou and T. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 6799-6808, 2017.

[71] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE computer society conference on computer vision and pattern recognition (CVPR),* Vol. 1, pages 886-893, 2005.

[72] B.J. Prosser, W.S. Zheng, S. Gong, T. Xiang and Q. Mary. Person re-identification by support vector ranking. In *Proceedings of the British Machine Vision Conference (BMVC),* 2010.

[73] W.S. Zheng, S. Gong and T. Xiang. Re-identification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence,* 35(3):653-668, 2013.

[74] P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence (TPAMI),* 19(7):711-720, 1997.

[75] A.R. Webb. Statistical pattern recognition. *John Wiley & Sons,* 2003.

[76] T. Ojala, M. Pietikäinen and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition,* 29(1):51-59, 1996.

[77] Y. Zhang and S. Li. Gabor-LBP based region covariance descriptor for person re-identification. In *Sixth International Conference on Image and Graphics,* pages 368-371, 2011.

[78] Y. Zhang, B. Li, H. Lu, A. Irie and X. Ruan. Sample-specific svm learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 1278-1287, 2016.

[79] L. An, M. Kafai, S. Yang and B. Bhanu. Reference-based person re-identification. In *10th IEEE international conference on advanced video and signal based surveillance,* pages 244-249, 2013.

[80] L. Wei, S. Zhang, W. Gao and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 79-88, 2018.

[81] J. Almazan, B. Gajic, N. Murray and D. Larlus. Re-id done right: towards good practices for person re-identification. *arXiv preprint arXiv:1801.05339,* 2018.

[82] A. Paszke, S. Gross, S. Chintala and G. Chanan. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration, 2017. Available: https://pytorch.org/

[83] F. Hafner, A. Bhuiyan, J.F. Kooij and E. Granger. A cross-modal distillation network for person re-identification in rgb-depth. *arXiv preprint arXiv:1810.11641*, 2018.

[84] V. Kumar, A. Namboodiri, M. Paluri and C.V. Jawahar. Pose-aware person recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 6223-6232, 2017.

[85] C. Su, J. Li, S. Zhang, J. Xing, W. Gao and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE international conference on computer vision (CVPR),* pages 3960-3969, 2017.

[86] R. Zhao, W. Oyang, X. Wang. Person re-identification by saliency learning. *IEEE transactions on pattern analysis and machine intelligence,* 39(2):356-370, 2016.

[87] W. Li, R. Zhao, T. Xiao, X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 152-159, 2014.

[88] X. Wang, R. Zhao. Person re-identification: System design and evaluation overview. In *S. Gong, M. Cristani, S. Yan, C. Loy (eds) Person Re-Identification. Advances in Computer Vision and Pattern Recognition,* pages 351-370, Springer, London, 2014.

[89] L. Wu, C. Shen, A.V.D. Henge., Personnet: Person re-identification with deep convolutional neural networks. *arXiv pre-print arXiv:1601.07255,* 2016.

[90] L. Zhang, T. Xiang, S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR),* pages 1239-1248, 2016.

[91] M. K. Uddin, A. Lam, H. Fukuda, Y. Kobayashi, Y. Kuno. Depth Guided Attention for Person Re-identification. In *DS. Huang, P. Premaratne (eds) Intelligent Computing Methodologies, ICIC*, vol. 12465, pages 110-120, Springer, 2020.

[92] K.Q. Weinberger, L.K. Sau. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research,* 10(2), 2009.

[93] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg. Imagenet large scale visual recognition challenge. *International journal of computer vision,* 115(3):211-252, 2015.

[94] Z. Imani, H. Soltanizadeh, A.A. Orouji. Short-Term Person Re-identification Using RGB, Depth and Skeleton Infor-mation of RGB-D Sensors. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering,* 44:669–681, 2020.

[95] N. Karianakis, Z. Liu, Y. Chen, S. Soatto. Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In *European conference on computer vision (ECCV),* pages 715-733, 2018.

[96] Z. Yu, Y. Zhao, B. Hong, Z. Jin, J. Huang, D. Cai, X. He, X.S. Hua. Apparel-invariant Feature Learning for Apparel-changed Person Re-identification. *arXiv preprint arXiv:2008.06181*, 2020.

[97] L. Cruz, D. Lucio, L. Velho. Kinect and RGBD Images: Challenges and Applications. In *25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials,* pages 36-49, 2012.