PART-BASED PERSON RE-IDENTIFICATION IN A CAFE SETTING

by

Hilal Özdemir Pehlivan

B.S., Computer Engineering, Boğaziçi University, 2015

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2018

PART-BASED PERSON RE-IDENTIFICATION IN A CAFE SETTING

APPROVED BY:

Prof. Lale Akarun .....
(Thesis Supervisor)

Assist. Prof. Emre Uğur .....

Assist. Prof. M. Furkan Kıraç .....

DATE OF APPROVAL: 25.07.2018

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my research supervisor Professor Lale Akarun for her valuable guidance, encouragement and support.

I also would like to thank my thesis committe members Assist. Prof. Emre Uğur and Assist. Prof. M. Furkan Kıraç for helping me improve my thesis with their valuable feedback.

I would like to thank my teachers in Keio University, Hyper Vision Research Laboratory (HVRL) Prof. Saito Hideo and Assist. Prof. Ozasa Yuko for accepting me to HVRL and always supporting me on my research.

I also would like to thank members of HVRL, Hachiuma Ryo, Ienaga Naoto, Paulo Dores, Yamaguchi Masahiro, Nakashima Masaki, Kikuchi Toshiki and Oniki Kazumasa for their friendship, kindness and generous help.

Lastly, I would like to thank my husband for his continuous encouragement and endless support.

This work was partially supported by JST CREST Intelligent Information Processing Systems Creating Co-Experience Knowledge and Wisdom with Human-Machine Harmonious Collaboration.

ABSTRACT

PART-BASED PERSON RE-IDENTIFICATION IN A CAFE SETTING

Person re-identification (re-id) is a challenging task particularly when people have various poses and some parts of the person's body are occluded or missing in the view. In this thesis, we propose a part-based person re-id method to cope with these challenges. The proposed method does not require any alignment in input images and is able to work even if only a part of the body is seen. In the proposed method, with the help of a pose estimator, body parts are detected and extracted from the input images. The extracted body parts are processed separately through associated deep learning models. Given two input body part images, each model outputs the probability that the given body parts extracted from two different raw images belong to the same person. In the fusion step, outputs of each body part model are combined and the result is output as the final probability value. We investigated three fusion methods and empirically showed that averaging the available body part model outputs is the best fusion method among the three. In order to develop and evaluate our proposed method, we collected the Robot Cafe dataset which abounds with the challenges mentioned above. For this purpose, we developed an annotation tool to easily and fastly select and extract person images from videos. Robot Cafe dataset has 10,969 images from 93 persons. The experiments are conducted on Robot Cafe and CUHK03 datasets. The experiment results showed that our method is more robust to missing body parts and huge pose changes compared to some of the previous studies.

ÖZET

KAFE ORTAMINDA KISMA DAYALI KİŞİYİ YENİDEN TANIMA

Kişiyi yeniden tanıma, kişilerin pozlarının çeşitlilik gösterdiği ve vücutlarının bir kısmının görüntü içinde olmadığı durumlarda oldukça zorlu bir problemdir. Bu tezde, bu tip zorlu durumlarda da çalışabilecek, kısma dayalı bir kişiyi yeniden tanıma yöntemi önerilmektedir. Önerilen yöntem, girdi imajlarının hizalı olmasını gerektirmemekte ve kişi vücudunun yalnıca bir kısmının görülebildiği durumlarda da çalışabilmektedir. Önerilen yöntemde, bir poz tahmin edici yardımıyla, girdi imajlardaki vücut kısımları saptanır ve çıkarılır. Çıkarılan vücut kısımları, ilgili derin öğrenme modelleri ile ayrı ayrı işlenir. Her model, verilen iki vücut kısmı imaj girdisi için, iki farklı ham imajdan çıkarılan bu vücut kısımlarının aynı kişiye ait olma olasılıklarını çıktı verir. Füzyon aşamasında, vücut kısmı modellerinin çıktıları birleştirilir ve son olasılık değeri olarak çıktı verilir. Üç füzyon yöntemi incelenmiş ve üçü arasında en iyi yöntemin, mevcut vücut kısmı modeli çıktılarının ortalamasının alınması olduğu deneysel olarak gösterilmiştir. Yöntemin geliştirilmesi ve değerlendirilmesi amacıyla, Robot Kafe veri kümesi toplanmıştır. Robot Kafe veri kümesi, içinde yukarıda değinilen zorlukları barındırmaktadır. Veri kümesinin hazırlanması için bir işaretleme aracı geliştirilmiştir. Bu işaretleme aracı, videolardan kişi imajlarının kolay ve hızlı bir şekilde çıkarılmasını sağlamaktadır. Robot Kafe veri kümesi, 93 kişiye ait 10,969 imaj içermektedir. Deneyler, Robot Kafe ve CUHK03 veri kümeleri ile gerçekleştirilmiştir. Deney sonuçları, önerilen yöntemin eksik vücut kısımları ve poz çeşitliliği gibi zorluklara karşı varolan bazı yöntemlerden daha güçlü olduğunu göstermiştir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . viii
LIST OF TABLES . . . xi
LIST OF SYMBOLS . . . xii
LIST OF ACRONYMS/ABBREVIATIONS . . . xiii
1. INTRODUCTION . . . 1
2. RELATED WORK . . . 4
2.1. Person Re-Identification . . . 4
2.1.1. Performance Measures . . . 9
2.1.2. Person Re-Identification with Color Cameras . . . 9
2.1.3. Person Re-Identification with RGB-D Cameras . . . 13
2.1.4. Person Re-Identification with Active Vision . . . 14
2.2. Datasets . . . 14
3. ROBOT CAFE DATASET . . . 21
3.1. Dataset . . . 21
3.2. Annotation Tool . . . 22
3.2.1. Frame Selection . . . 23
3.2.2. Pose Estimation . . . 23
3.2.3. Person Image Selection . . . 24
3.2.4. Person ID Annotation . . . 24
4. PERSON RE-IDENTIFICATION BY DEEP PART LEARNING . . . 27
4.1. Extraction of Body Parts . . . 29
4.2. Network Architecture . . . 29
4.3. Fusion of Individual Model Results . . . 31
4.3.1. Fusion Method 1 . . . 31
4.3.2. Fusion Method 2 . . . 31
4.3.3. Fusion Method 3 . . . 32

4.4. Training 33
4.5. Fine-tuning 34
5. EXPERIMENTS AND RESULTS 35
5.1. Comparison of Fusion Methods 35
5.2. Experiments on Robot Cafe Dataset 36
5.3. Experiments on CUHK03 Dataset 40
6. CONCLUSION 44
REFERENCES 45

LIST OF FIGURES

Figure 2.1. Example views from CUHK03 database [1]. 14

Figure 2.2. Example views from Market-1501 dataset [2]. 15

Figure 2.3. Example views from VIPeR database [3]. 15

Figure 2.4. Example views from i-LIDS database [4]. 16

Figure 2.5. Example views from ETHZ dataset [5]. 17

Figure 2.6. Example views from database created and used by [6]. 17

Figure 2.7. Example views from database created and used by [7]. 17

Figure 2.8. Example views from database created and used by [8]. 18

Figure 2.9. Example views from MRP Drone datasets [9]. 18

Figure 3.1. Example images from Robot Cafe dataset [10]. 21

Figure 3.2. The cafe layout and camera setting. 22

Figure 3.3. The annotation process. 23

Figure 3.4. Keypoints detected by OpenPose [11]. 23

Figure 3.5. How to use keys to navigate. 24

Figure 3.6. A screen shot of the annotation process. 25

Figure 3.7. Data folder structure of each person. 26

Figure 4.1. Flow of the system. 28

Figure 4.2. Example of an input image cropped into available body parts. . . 29

Figure 4.3. Proposed architecture which is a modified version of the architecture of Ahmed *et al.* [12]. 30

Figure 4.4. Fusion method 2. An additional fully-connected layer is used to fuse the individual model outputs. 32

Figure 4.5. Fusion method 3. Individual model outputs are weighted according to the keypoint detection confidence values and then averaged to fuse the individual model outputs. 33

Figure 5.1. Training process of the fusion layer in the fusion method 2. 35

Figure 5.2. Comparison of investigated fusion methods. 35

Figure 5.3. CMC curves on Robot Cafe dataset comparing our method to Improved Deep [12] after fine-tuning. 36

Figure 5.4. Fine-tuning process of each body part model for camera pair a-b. . 37

Figure 5.5. Fine-tuning process of each body part model for camera pair b-c. . 38

Figure 5.6. Fine-tuning process of each body part model for camera pair a-c. . 39

Figure 5.7. CMC curve on CUHK03 dataset comparing our method to some of the previous studies. 40

Figure 5.8. Detection percentages of keypoints in CUHK03 dataset. 41

Figure 5.9. Detection percentages of keypoints in Robot Cafe dataset. 41

Figure 5.10. Number of body parts used in comparisons. 41

Figure 5.11. Training process of each body part model. 42

Figure 5.12. CMC curves of individual body part models on CUHK03 dataset. 43

LIST OF TABLES

Table 2.1. Survey of person re-id studies. 6

Table 2.2. Survey of person re-id datasets. 19

Table 3.1. How the data is split in Robot Cafe dataset [10]. 21

LIST OF SYMBOLS

A	Input image
A_k	The kth body part in the image A
B	Input image
B_k	The kth body part in the image B
$\mathcal{I}$	The set of the indexes of available body parts
K	The number of body parts available in both image A and image B
n	Learning rate
p_k	Output of the kth model
λ	Weight decay
μ	Momentum

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
AI	Artificial Intelligence
CMC	Cumulative Matching Characteristic
CNN	Convolutional Neural Network
ELF	Ensemble of Localized Features
FN	False Negative
FP	False Positive
GAP	Global Average Pooling
HSV	Hue Saturation Value
MRP	Mobile Re-identification Platform
MSCR	Maximally Stable Color Regions
MvsM	Multiple versus Multiple
MvsS	Multiple versus Single
PLS	Partial Least Squares
RGB	Red Green Blue
RGB-D	Red Green Blue - Depth
RHSP	Recurrent High-Structured Patches
SDALF	Symmetry-Driven Accumulation of Local Features
SRR	Synthetic Recognition Rate
SvsS	Single versus Single
TN	True Negative
TP	True Positive

1. INTRODUCTION

With the developments in robot technology and Artificial Intelligence (AI) research, robots are expected to assume more roles in our daily lives in the near future. Robot cafe refers to a cafe concept where the waiter or the waitress is replaced with a robot. In such context, a robot waiter or waitress would be expected to welcome customers in the entrance, guide them to an appropriate table according to the number of customers, take the customer orders after customers sit down, respond to customers when called, serve orders correctly to the customers who ordered them, receive payments, communicate with customers, etc.

In order for robots to achieve all of these tasks successfully, the necessary abilities include person detection, natural language processing, and action recognition. Also, in order for robots to serve customers correctly, it is necessary for robots to recognize an individual when they see him/her again. For instance, when serving an order, the robot should be able to find the customer who ordered it. Similarly, when receiving a payment, the robot should be able to understand who the customer is so that it can remember the customer's orders and receive the payment accordingly. Therefore, we aim to achieve this task by applying person re-identification technique in robot vision.

Person re-identification (re-id) is a technique for recognizing an individual in non-overlapping views of multiple cameras. In order to achieve person re-id task, the most popular approach is to use the visual information of individuals. The difference between biometric verification and person re-identification is that biometric verification should work even if the general appearance of a person changes. On the other hand, person re-identification assumes that persons' appearances are constant. If a person's clothing differs between cameras, person re-id system could not match the person correctly. On the other hand, biometric verification methods require more constraints, such as a clear, frontal face image with no enhanced emotions, or a fingerprint, or an iris image, or a signature to be taken; whereas person re-id techniques are more flexible and feasibly applicable without forcing persons to change their natural behaviors. In

cafe setting, using face verification techniques may not be feasible, since a clear image of the customer's face may not always be available. However, person re-id techniques can be feasible applied, since a general image of a person expected to be sufficient to verify his/her identity.

In recent years, person re-id has gained a considerable attention. New techniques, as well as new datasets have been introduced. Unlike the old studies using hand-crafted features, studies using deep learning techniques has become popular. To our knowledge, there has not been any publication mentioning an existing person re-id application. However, in various studies it is mentioned that possible applications of person re-id might be in surveillance, i.e. for fastly searching for a lost child or a fugitive person in a network of surveillance cameras, or in marketing for understanding customer behaviors during shopping. Even if current state-of-the-art person re-id methods cannot always successfully match persons in their first match, they mostly achieve to find the correct match in the first 10 to 20 matches. Therefore, existing person re-id methods can make person re-id by humans faster and more feasible by reducing the number of candidate matches to select from.

There are several RGB image and video databases which have been used for evaluation of person re-identification systems. These databases are built either with views of two cameras, or using a network of surveillance cameras, or using hand-held cameras, or using a camera attached to a drone. There are also RGB-D databases which are collected indoors for person re-identification task.

In this thesis, we propose a person re-identification method that is able to meet the challenges in a cafe setting: Partially occluded persons, sitting persons, and partially visible bodies. Our method is based on deep part-based learning. Each body part is processed through a separate deep learning model, therefore our method does not put any constraint on human body pose. Also, our fusion method does not require all body parts to be available in the image, therefore our method is also applicable when there are occlusions or missing body parts.

Most of the person re-id datasets consist of overall, standing images of persons. However, in the cafe context, persons may be standing or sitting and sometimes only a part of their body may be available in camera view. In order to develop a system to achieve person re-id task in the cafe context, existing datasets would not be sufficient. Therefore, as a part of this study, we also collected a new dataset for person re-identification in cafe setting.

The rest of the thesis is organized as follows: Chapter 2 reviews previous work in person re-identification. Chapter 3 introduces the Robot Cafe dataset and the annotation tool. Chapter 4 describes the technique developed, person re-identification by deep part learning. Chapter 5 gives the experiments and results. Finally, Chapter 6 present the conclusions.

2. RELATED WORK

2.1. Person Re-Identification

In person re-identification, the aim is to match the detected human views belonging to the same individual in a multiple camera setting. If the camera views are overlapping and/or there is information about the camera positions in the environment, applying temporal reasoning to the spatial data is an option to solve this problem. However, this is not always the case, so appearance-based person re-identification systems allow us to match individuals with the features extracted from their individual appearances, without having to know an extra information about camera positions in the environment.

Appearance-based person re-id can be difficult to achieve in the occurrence of the following cases: changes in illumination, changes in person body pose, different camera angles, occlusions, missing body parts, different persons with similar clothing, different front and rear appearances.

Most of the person re-id studies assume that person detection is applied to raw images beforehand. Their techniques are applied to already detected person images and their evaluation criteria do not include the success of the person detection part. Therefore, we can say that person detection is a prerequisite for person re-identification.

In person re-id studies, the test set in the preferred person re-id dataset is divided into two sets as gallery and probe. Both sets have images from the same set of person identities from different cameras. The aim is to match the identities in the probe set to known identities in the gallery set. Depending on the dataset, the probe and gallery sets may have single or multiple images for each identity. When the gallery and probe sets include only one image per person, experiments can be conducted as Single versus Single (SvsS) comparisons, meaning that person images in the probe set are matched to a person in the gallery set by comparing them to only one image for each identity.

When there are multiple images in the gallery and probe sets, experiments can be conducted as SvsS, Multiple versus Multiple (MvsM) or Multiple versus Single (MvsS) comparisons. In MvsM, multiple images of each person in the probe set are together compared to multiple images of each person in the gallery set. In MvsS, a single image of each person in the probe set is compared to multiple images of each person in the gallery set.

The difference between person tracking and person re-id is that person tracking is applied within camera where a person is observed in consecutive frames until he/she leaves the view. With the help of homographic transformation, person tracking can also be applied across cameras with overlapping views [13]. On the other hand, person re-id can be applied both within camera even if persons leave from and enter into the view, and across cameras with overlapping or non-overlapping views. In some of the person re-id studies, person tracking is used to obtain person templates which are defined as the set of detected person's images in a pre-defined number of consecutive frames. Then person re-id is applied between these templates [8].

In Table 2.1, an overview of person re-id studies are presented. They are explained in detail in the following subsections.

Table 2.1. Survey of person re-id studies.

	Year	**Problem Domain**	**General Approach**	**Features**	**Classifier/ Distance Measure**	**Dataset**	**Performance (Rank-1 re-id rate)**
[14]	2008	color camera	feature-based	color, texture	a set of likelihood ratio tests	VIPeR [3]	12%
[5]	2009	color camera	feature-based	color, texture, edge	Euclidean distance	ETHZ [15]	0.8%, 0,73% and 0.77% for sequences #1, #2 and #3, respectively
[16]	2010	color camera	feature and part-based	color histograms, MSCR and RHSP	SDALF [16] matching distance	VIPeR [3], iLIDS [17], ETHZ [15]	20%, 40%, 0.90%, 0.91% and 0.94% for sequences #1, #2 and #3, respectively

Table 2.1. Survey of person re-id studies (cont.).

[18]	2017	color camera	CNN and part-based	CNN features	Euclidean distance	Market-1501 [2], CUHK03 [1], VIPeR [3]	89%, 85% and 82% for labeled and detected sets, respectively, 54%
[19]	2017	color camera	CNN	CNN features	Euclidean distance	Market-1501 [2], CUHK03 [1]	94%, 96.1%
[20]	2017	color camera	CNN and part-based	CNN features	classifier in CNN	Market-1501 [2], CUHK03 [1], VIPeR [3]	69.3%, 82.75%, 56.65%
[21]	2017	color camera	CNN and part-based	CNN features	Softmax classifier	VIPeR [3], CUHK Campus [22]	30.16%, 28.45%
[23]	2017	color camera	CNN and part-based	CNN features	Euclidean distance	Market-1501 [2], CUHK03 [1], VIPeR [3]	77.97%, 67.10%, 54.49%
[7]	2012	RGB-D camera	feature-based	3D soft biometric features	normalized squared distance	an RGB-D dataset [7]	15%

Table 2.1. Survey of person re-id studies (cont.).

[6]	2012	RGB-D camera	feature-based	color	a normalized weighted correlation coefficient	an RGB-D dataset [6]	93% for front-front pairs, 55% for front-rear pairs
[9]	2014	active color camera	feature-based	color, texture	Nearest Neighbor	MRP Drone datasets [9]	0.83 and 0.34 precision for intra flight and inter flight tests, respectively

2.1.1. Performance Measures

In person re-identification, Cumulative Matching Characteristic (CMC) curve and Synthetic Recognition Rate (SRR) curve are usually used as performance measures. CMC curve shows the probability of finding the correct match in the first n matches. SRR curve shows the probability of finding the correct match in the first rank, for n targets. The rank-1 person re-id rate is also mentioned in most of the studies, which is the first value plotted in a CMC curve.

In R. Layne *et al.* [9]'s study, due to the different nature of re-identification with moving cameras, three problem variants and metrics are defined. When there is a pre-defined watchlist and the task is to understand whether an encountered person is on the watchlist or not, the rank of the true matches and average precision are used. For within-flight and across-flight re-identification, precision, recall, f-measure, specificity and accuracy are used (see Eq. 2.1).

$$\begin{aligned} Precision &= TP/(TP+FP) \\ Recall &= TP/(TP+FN) \\ Specificity &= TN/(FP+TN) \\ Accuracy &= (TP+TN)/N \\ F\text{-}measure &= 2 * \frac{precision * recall}{precision + recall} \end{aligned} \tag{2.1}$$

2.1.2. Person Re-Identification with Color Cameras

Gray and Tao [14] introduce a representation called an ensemble of localized features (ELF). The idea is, instead of using a hand-crafted feature, to define a set of features and similarity measures, then to use a machine learning algorithm to select the final model with maximum discriminative power for a given training data. Investigated

feature channels are eight color channels from RGB, YCbCr, and HSV colorspaces and nineteen texture channels which are Schmid [24] and Gabor [25] filters constructed with various parameters. As the similarity measure, a set of likelihood ratio tests with Exponential, Gamma, and Gaussian distributions are investigated. They conduct experiments on VIPeR dataset [3] and report the rank-1 person re-id accuracy as 12%. They also provide the plot of re-id rate versus the number of possible targets up to 25. When the number of targets is 25, the re-identification rate is reported to be 50

Schwartz *et al.* [5] first create a very high dimensional feature space with color, texture and edge features, then used Partial Least Squares (PLS) technique to reduce the dimensionality. For each person, a different PLS discriminatory model is learned. During testing, a test sample's feature vector is processed with PLS models belonging to each person in the gallery set and Euclidean distances are calculated. Then, it is assigned to the identity of the person with the smallest distance value. They conducted experiments on ETHZ dataset [15]. Rank-1 person re-id accuracies are reported to be 0.8, 0.73, 0.77 for ETHZ dataset's sequences #1,#2 and #3, respectively.

In Farenzena *et al.* [16]'s technique, which they called Symmetry-Driven Accumulation of Local Features (SDALF), they first determine asymmetry axis and divide the body into two part as torso and legs and then determine symmetry axis for each. Then from each part, three features are extracted: weighted color histograms in HSV space, maximally stable color regions (MSCR) and recurrent high-structured patches (RHSP). Finally, these features are jointly used as a single signature during matching. A unique distance measure called the SDALF matching distance is used for matching. Experiments are conducted on VIPeR [3], iLIDS [17] and ETHZ [15] datasets. They report rank-1 person re-id rate on VIPeR dataset to be around 20%. For iLIDS and ETHZ datasets, since there are multiple images of each person in gallery and probe sets, SvsS, MvsS and MvsM tests are conducted. Rank-1 person re-id rate in MvsM comparisons on iLIDS dataset is 40% and on ETHZ dataset is 90%, 91% and 94% for sequences #1, #2 and #3, respectively.

Wei *et al.* [18] use a partial representation, where they only use body parts that can be detected easily and reliably: the head, the upper body, and the lower body. Detected body parts and the original image is given to 4 sub-networks which have the same architecture and weights. The reason behind weight sharing is that this way more generic features are learned instead of paying too much attention to the details, which may not be a good idea when we have misalignment and pose changes. Outputs of CNNs are given to two consecutive convolutional layers and Global Average Pooling (GAP) is applied to the resulting feature maps in order to generate the final classification scores, called descriptors. GAP, as the name suggests, outputs the average of values in a feature map. At the end of the descriptor learning process, four descriptors are concatenated to get a final descriptor. During matching, first, gallery set is clustered into groups and a single template descriptor for each group is generated. When a test image is given, its descriptor is calculated and then it is compared to all template descriptors. Images in the relevant groups are retrieved and compared to the test image. Then the final classification decision is made. This process is preferred to reduce the retrieval time. During comparisons, Euclidean distance between descriptors is used. Experiments are conducted on Market-1501 [2], CUHK03 [1] and VIPeR [3] datasets. Rank-1 person re-id rates are reported as 89%, 85%, 82% and 54.8% on Market-1501, CUHK03 (labeled), CUHK03 (detected) and VIPeR datasets, respectively.

Zhang *et al.* [19] first divide images into horizontal strips and then match the strips in two images in a way that will result in minimum total distance between matched strips. They use a CNN to calculate feature maps. Global pooling is applied on the obtained feature maps to get global features and global pooling in the horizontal direction is applied to get local features. L2 distance between global features and L2 distances between matched local features are summed and used as the distance measure between two images. Experiments are conducted on Market-1501 [2] and CUHK03 [1] datasets. Rank-1 person re-id rates are reported as 94% and 96.1% on Market-1501 and CUHK03 (labeled) datasets, respectively.

Yao *et al.* [20] start with producing feature maps by giving person images to their network trained for classification. Then, the produced feature maps are clustered into

four groups. Each feature map group is average pooled and resulting maps show the most salient body parts which turn out to be head, upper body, lower body, and feet. Then they continue with computing the loss for each part. This process is repeated in each iteration of the network training process. During testing, test images are given as the input to the pre-trained classification network to get the feature maps. Then using all feature maps, a global representation is obtained. Part representations are obtained from the trained part loss networks. Global and part representations are concatenated as the final representation and used for similarity computation. Experiments are conducted on Market-1501 [2], CUHK03 [1] and VIPeR [3] datasets. Rank-1 person re-id rates are reported as 69.3%, 82.75% and 56.65% on Market-1501, CUHK03 and VIPeR datasets, respectively.

Zhao *et al.* [21] start with body region extraction. They trained two separate networks for feature extraction and feature fusion. Inside of the feature extraction network, each body part and the whole body are given to separate CNNs. Feature vectors obtained from the feature extraction network are combined in the feature fusion network and the final feature vector is produced. Softmax function is used to calculate the network loss. How they handled the case of undetected or missing body parts is not mentioned in the paper. Experiments are conducted on VIPeR [3] and CUHK Campus [22] datasets. Rank-1 person re-id rates are reported as 30.16% and 28.45% on VIPeR and CUHK Campus datasets.

Zheng *et al.* [23] start with detecting body parts and then construct the *PoseBox*, which combines the detected body parts in a certain way to prevent misalignments during matching of detected bodies in two images. They input the whole body and the constructed *PoseBox* to two CNNs having the same architecture with different weights. They solve the problem of a possible poor pose estimation by inputting the pose estimation confidence vector to their fusion network so that the reliability of the *PoseBox* can be assessed and the original image may have a bigger role in the extracted descriptor. The whole body, the *PoseBox* and pose estimation confidence vector are combined in the fully connected layers. The Euclidean distance between fully-connected layer activations is used during testing as distance measure. Experiments are conducted

on Market-1501 [2], CUHK03 [1] and VIPeR [3] datasets. Rank-1 person re-id rates are reported as 77.97%, 67.10% and 54.49% on Market-1501, CUHK03 and VIPeR datasets, respectively.

2.1.3. Person Re-Identification with RGB-D Cameras

Barbosa *et al.* [7] presents a set of 3D soft-biometric cues gathered using an RGB-D sensor in order to use them as features in person re-identification task. The advantage of using these features is that even if persons change their clothing, soft-biometric cues mostly remains unaffected. Normalized squared differences between features are calculated and their weighted sum is used as the distance measure. In order to evaluate their method, they collected an indoor RGB-D dataset with four groups of data. Experiments are conducted on several combinations of these data groups being probe and gallery sets. The rank-1 person re-id rate in one combination is reported as 15%.

Albiol *et al.* [6] also attempts to solve person re-identification problem using Kinect cameras. In the proposed method, both RGB and depth cameras in Kinect are calibrated and height maps are created for each frame. A height map refers to the image where the pixel values represent the height with respect to the ground. People are segmented using thresholding and connected component analysis on height maps. These blobs are tracked for several frames. From these tracks, bodyprints are obtained for each person by dividing the body into height levels and summarizing the color appearance for each level. These bodyprints then are used to match different tracks belonging to the same person. During matching, a normalized weighted correlation coefficient is proposed and used. This measure takes the confidence of detected height level into account so that the system is robust to occlusions or undetected height levels. For the evaluation of the system, when frontal views and rear views are handled separately, the average correct re-identification performance is 93%, and when frontal and rear views are compared, result drops to 55%. Both studies are only applicable for indoor use, because of the Kinect's technology.

2.1.4. Person Re-Identification with Active Vision

R. Layne *et al.* [9] introduces the first dataset for mobile person re-identification. Unique challenges of mobile re-identification are addressed. Due to the different nature of open-world mobile person re-identification, information retrieval style metrics are used to evaluate the success of several algorithms. The ELF [14] features of detected persons are extracted and used. It is concluded that in general simpler methods, such as Nearest Neighbor matching, outperforms more complex methods in Mobile Re-Identification Platform (MRP) re-id. The precision of the system is reported as 0.83 for intra-flight tests and 0.34 for inter-flight tests.

2.2. Datasets

CUHK03 [1] dataset has 13,164 images of 1,360 persons captured by 6 cameras (see Figure 2.1). Each person has images from two of the cameras. The dataset is divided into training (1160 persons), validation (100 persons) and test (100 persons) sets. CUHK03 dataset has challenges such as misalignment, occlusions, and missing body parts. It is one of the largest person re-identification datasets in the literature and it has been used for training the deep learning models in several studies.



Figure 2.1. Example views from CUHK03 database [1].

Market-1501 [2] dataset is collected outdoors with six cameras. It has 32,668 images from 1,501 identities. There are also distractor images which are false positive or very poor detections of the person detector used. The dataset has challenges such as misalignment and missing body parts. (see Figure 2.2)



Figure 2.2. Example views from Market-1501 dataset [2].

VIPeR [3] is the most famous and challenging database for person re-identification studies. Pairs of 632 images of pedestrians are recorded (see Figure 2.3). Since only two images for each person is available, this database can only be used for evaluation of single-shot person re-identification systems. There is no occlusion in the images of this database. There are variations in pose, viewpoint, and lighting in person images.



Figure 2.3. Example views from VIPeR database [3].

The iLIDS-VID [4] database is created using the record of an airport arrival hall [26] with two non-overlapping camera views by cropping people from the original video. There are sequences of images belonging to 300 people (see Figure 2.4) from each camera. Sequence lengths variate between 23 to 192 image frames. In this database,

there are challenging situations such as occlusions, people with very familiar clothing, changes in lighting and cluttered background. This database is suitable for evaluation of both single-shot and multiple-shot person re-identification systems.



Figure 2.4. Example views from i-LIDS database [4].

ETHZ dataset [15] is created by using data collected from a mobile camera attached to a car. This dataset is mostly used in pedestrian detection and tracking studies. In Schwartz *et al.* [5]'s study, persons are cropped and then this new version of the dataset is used for evaluation of their appearance-based person recognition technique (see Figure 2.5). There are three sequences in the dataset. There are changes in pose and illumination in person images. There are 4,857 images from 83 persons, 1,936 images from 35 persons and 1,762 images from 28 persons in sequences #1, #2 and #3, respectively.

In Albiol *et al.* [6]'s study, an RGB-D dataset is collected for person re-identification with 40 people. Two Kinect cameras are located approximately at the same location but pointing at opposite directions. People are recorded as they come towards to the camera and as they go away from the camera, therefore there are front and rear views of people (see Figure 2.6). In this dataset, there are challenging situations such as occlusion, backpack use, long hair and opened jackets.

In Barbosa *et al.* [7]'s study, an RGB-D dataset is collected for person re-identification with 79 people. There are four groups of recordings. In "Collaborative" group, people are asked to walk with two arms stretched and with no occlusions. In second ("Walking1") and third ("Walking2") groups, frontal views are recorded. In fourth ("Backwards") group, rear views are recorded (see Figure 2.7).



Figure 2.5. Example views from ETHZ dataset [5].

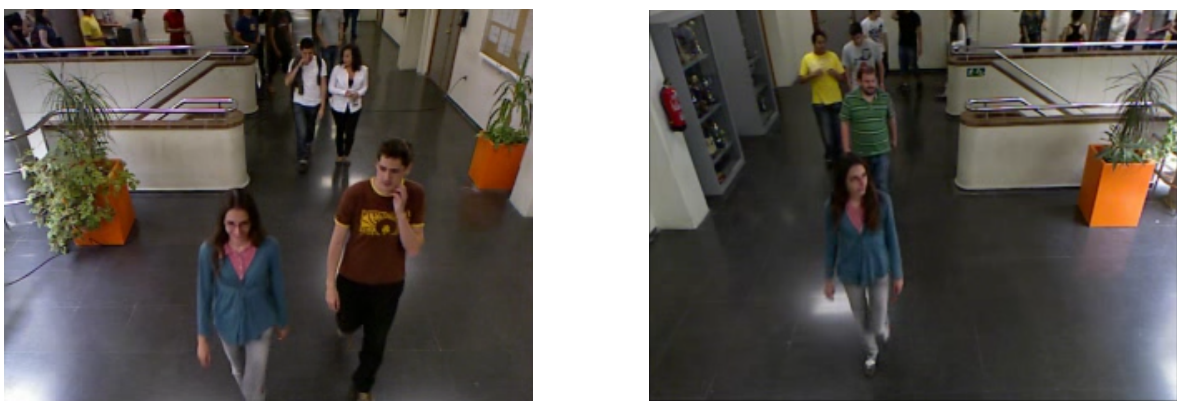


Figure 2.6. Example views from database created and used by [6].

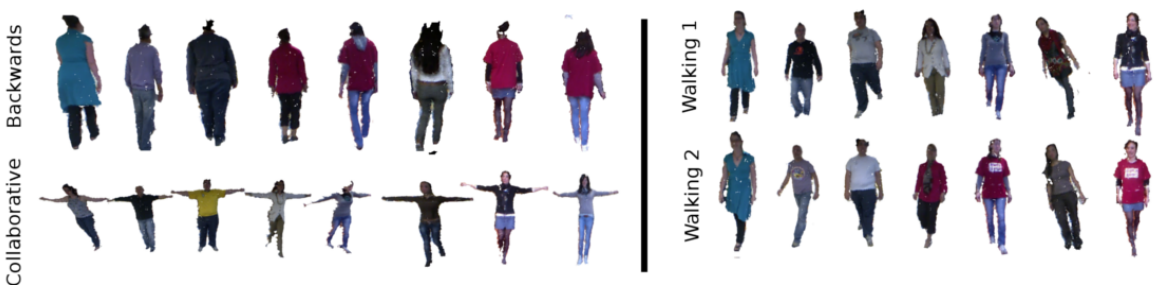


Figure 2.7. Example views from database created and used by [7].

In Satta *et al.* [8]'s study, an RGB-D dataset is collected for person re-identification with 54 people. There are frontal and back poses of individuals. In this database, there are challenging situations such as non-uniform illumination and changes in pose (see Figure 2.8).



Figure 2.8. Example views from database created and used by [8].

MRP Drone datasets [9] are first datasets for mobile person re-identification. Dataset 1 is collected with a total of three flights across indoor and outdoor environments, consists of 797 person detections from 10 distinct people in the longest flight. Dataset 2 is collected with a total of six flights in very crowded outdoor environments. In each flight, there are between 10000 and 30000 frames of video data and an average of 8654 person detections from an unknown number of distinct people (see Figure 2.9).



Figure 2.9. Example views from MRP Drone datasets [9].

An overview of person re-id datasets is presented in Table 2.2.

Table 2.2. Survey of person re-id datasets.

Database	**Year**	**Domain**	**# of Persons**	**# of Images**	**Challenges**
CUHK03 [1]	2014	outdoors	1,360	13,164	misalignment, occlusions, missing body parts
Market-1501 [2]	2015	outdoors	1,501	32,668	misalignment, missing body parts
VIPeR [3]	2007	pedestrian	632	1264	variations in pose, viewpoint and lighting
iLIDS-VID [26] [4]	1014	indoors	300	(avr.) 146/person	occlusions, similar clothing, variations in lighting, cluttered background
ETHZ [15] [5]	2007	pedestrian	83, 35, 28	4,857, 1,936, 1,762	variations in pose and lighting
An RGB-D dataset [6]	2012	indoors	40	video	occlusions, different views from front and back
An RGB-D dataset [7]	2012	indoors	79	video	changes in clothing
An RGB-D dataset [8]	2013	indoors	54	video	variations on pose and lighting
MRP Drone 1 [9]	2014	outdoors	10	797	variations in range, lighting, motion blur and alignment

Table 2.2. Survey of person re-id datasets (cont.).

Robot Cafe (Ours) [10]	2018	indoors	93	10,969	occlusions, missing body parts, similar clothing, different views from front and back, variations in pose and lighting

3. ROBOT CAFE DATASET

3.1. Dataset

Robot Cafe dataset [10] is collected indoors in a cafe setting. The dataset has 10,969 images from 93 persons captured by 3 cameras (a, b and c). Each person has up to 100 images from each camera. Some example images can be seen in Figure 3.1. Robot Cafe dataset is divided into training (53 persons), validation (20 persons) and test (20 persons) sets (see Table 3.1).



Figure 3.1. Example images from Robot Cafe dataset [10].

Table 3.1. How the data is split in Robot Cafe dataset [10].

Person ids	**Set**	**# of persons**
Person0 - Person52	training set	53
Person53 - Person72	validation set	20
Person73 - Person92	test set	20

Data is collected in six sessions in three days. People are recorded as they enter the cafe, give their orders, sit on a table, spend some time drinking and talking, then leave the cafe. The cafe layout and camera setting can be seen in Figure 3.2.

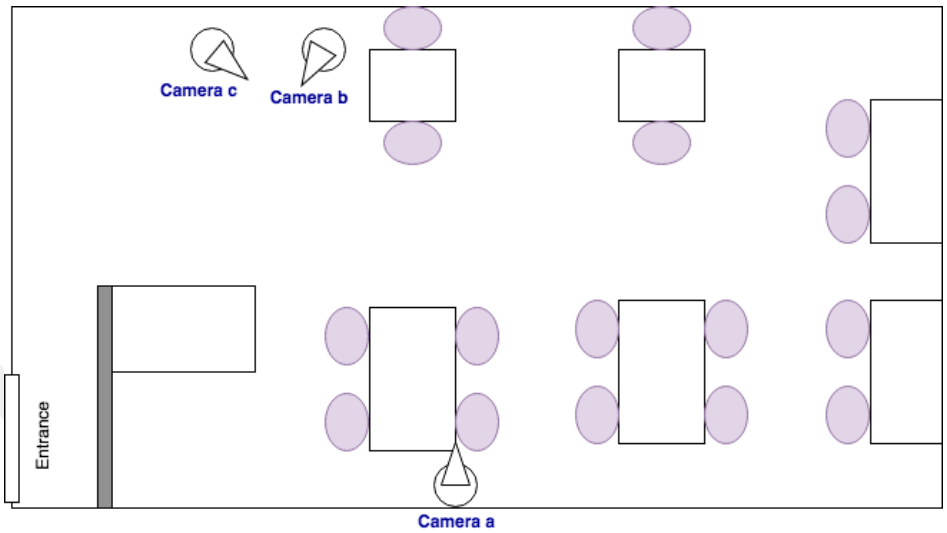


Figure 3.2. The cafe layout and camera setting.

The dataset includes challenges, such as occlusions, only some part of the body being available, very similar clothing, different views from front and back, variations in illumination, and huge pose changes where persons are sitting, standing, walking and bowing. Human detections are automatic, which enables us to apply our person re-id system in real scenarios.

3.2. Annotation Tool

In the process of creating the Robot Cafe dataset, in order to easily select and extract person images from videos, we developed an annotation tool. The annotation process consists of four steps: Frame selection, pose estimation, person image selection, and person id annotation. The annotation process is shown step by step in Figure 3.3. Steps which are shown in green are done by using the annotation tool.



Figure 3.3. The annotation process.

3.2.1. Frame Selection

The first step is to go over each video, frame by frame, and to select the frames that have desirable person images, by just clicking 'y' for 'yes' and 'n' for 'no'. After this step is completed, the selected frames are extracted as images.

The reason we select frames is that, if we give all of the frames in our videos to the pose estimator, it takes a very long time to complete. Selecting the useful frames at the beginning of the annotation process enables us to save on computation time.

3.2.2. Pose Estimation

As the second step, a pose estimation technique should be applied to the selected frames. We applied OpenPose [11] technique and obtained the detected keypoints information. Keypoints detected by OpenPose are shown in Figure 3.4. OpenPose technique is explained in more detail in Chapter 4.

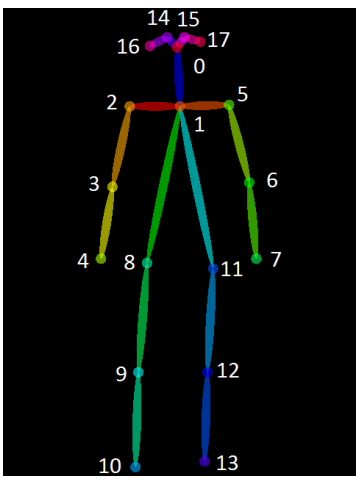


Figure 3.4. Keypoints detected by OpenPose [11].

3.2.3. Person Image Selection

The annotation tool allows the user to go over the frames which were selected in the first step and processed in the second step. Detected persons in a frame are first shown in red rectangles. If the user decides to add the person to the database, he/she presses 'y' for 'yes' and the rectangle becomes green. If a user changes his/her mind later, he/she can turn back to the person image and change the decision. In order to navigate between persons in a frame, 'w' and 's' keys can be used. In order to navigate between frames, 'a' and 'd' keys can be used (see Figure 3.5). It is okay to close the program any time during the process and when it is opened again, already annotated persons will be shown in the corresponding colors (see Figure 3.6).

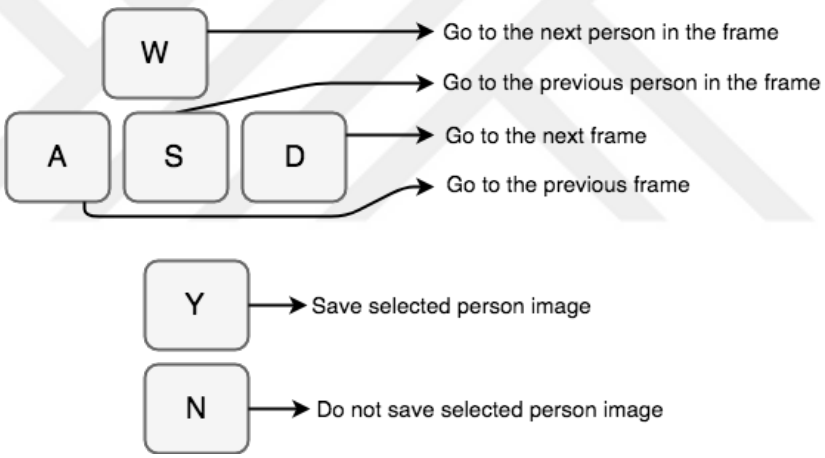


Figure 3.5. How to use keys to navigate.

3.2.4. Person ID Annotation

The last step is to group the selected person images according to identities by hand. Folder structure at the end of this procedure is shown in Figure 3.7. In each person's folder, there are three folders belonging to each camera. In a camera folder, there are images of the person cropped from videos. Each image has a unique name. There are also folders with the same name as their corresponding image file. In these folders, there are body part images cropped from raw images.

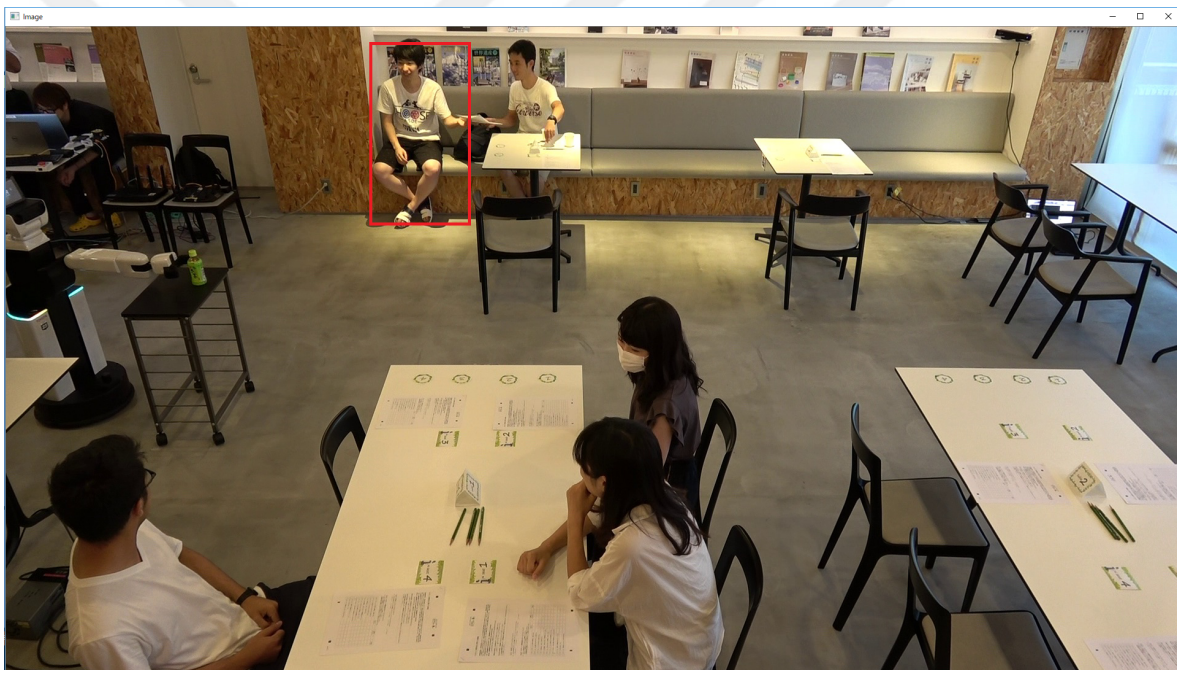


Figure 3.6. A screen shot of the annotation process.

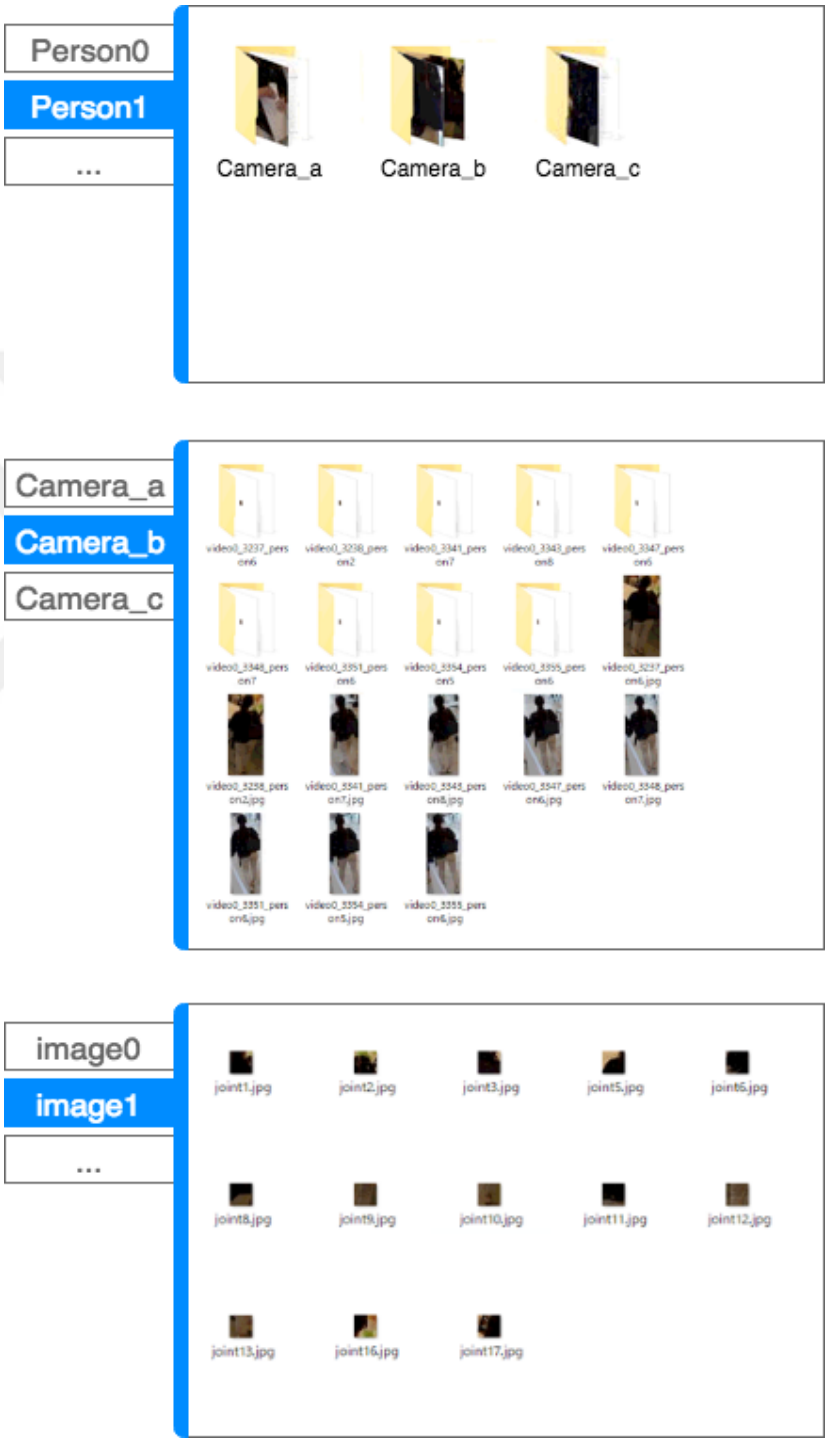


Figure 3.7. Data folder structure of each person.

4. PERSON RE-IDENTIFICATION BY DEEP PART LEARNING

Some of the person re-id studies have the assumption that if we segment person images into horizontal strips, strips having the same order will most of the time contain the same body parts [27] [28]. In reality, person body pose changes from image to image. Some studies vary the positional correspondence of image parts in order to solve this problem [12]. But in the case of large pose differences, such as a person standing in one image and sitting in another image, this approach would not be sufficient.

Another assumption most of the person re-identification studies make is that all person images has the whole body available [16]. This also may not hold in reality. For instance, one camera may capture the whole body whereas the other camera may have only a part of the body in its range. If we assume a positional correspondence in such case, we may end up with comparing, for instance, the legs in one image to the lower part of the torso in the other. In another case, an image may have a part of body occluded by an object or even by another person. If we compare it to another image in which the person's whole body is available, some parts of the body will be compared with the occluded parts of the body in the other image.

In this chapter, we introduce the use of individual body parts separately for person re-id task in order to cope with huge pose changes and missing body parts. First, body parts are extracted from the input image. Second, for each body part, a separate deep learning model is trained. Models take an image pair as input and output the probability that the two images belong to the same person. Individual part model probabilities are fused and output as the final probability value. See Figure 4.1.

Convolutional Neural Network (CNN) architecture that we used is inspired by the architecture of Ahmed *et al.* [12]. Their network is designed to tackle person re-id problem. The network takes two 160×60 person images and returns the probability

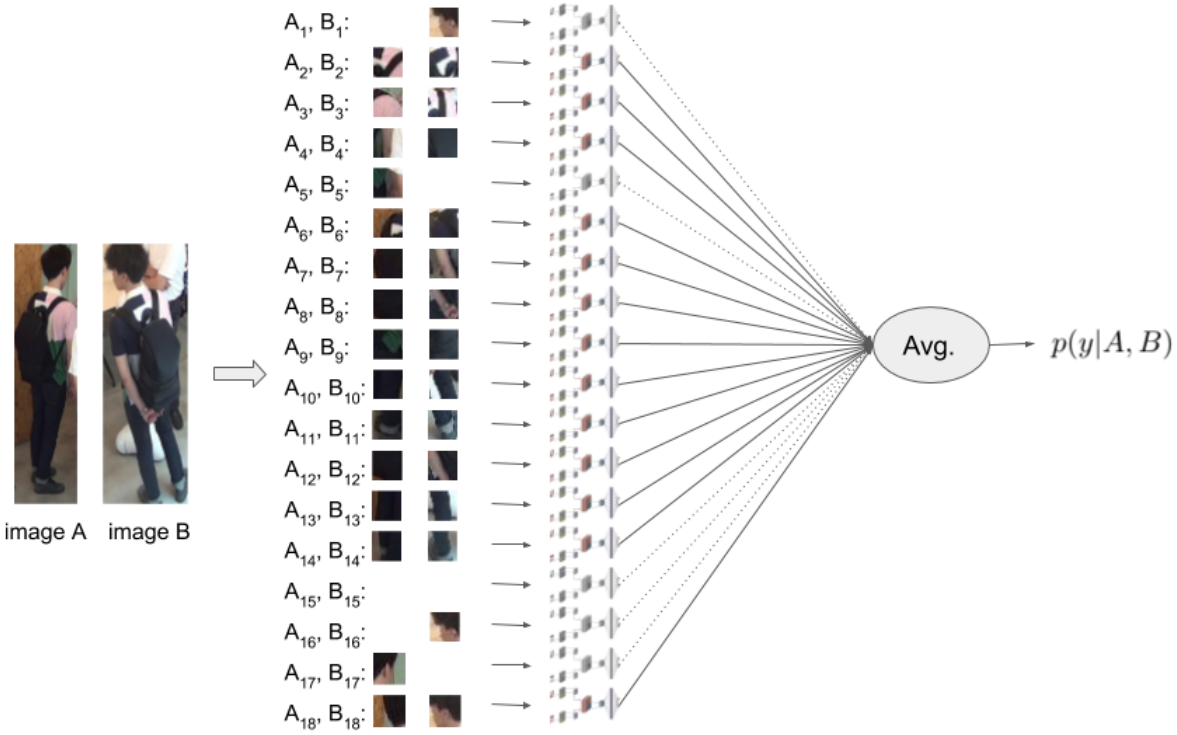


Figure 4.1. Flow of the system.

that two images belong to the same person. First, two images are separately fed into a convolutional layer, then the resulting feature maps are again fed into another convolutional layer. The reason why the images are processed with the same weights in the first two layers is that by doing so, the same features are extracted from each image. After that, in order to compare these extracted features, the *cross-input neighborhood differences layer* is used. This layer takes two feature maps and outputs the comparison result as follows: Values of each position in one feature map is compared to values in a 5×5 rectangle centered at the same position, therefore 25 values are output in each comparison. This process is repeated for the other way around as well. The reason why the comparison is done within a neighborhood instead of a value by value comparison is that, this way any small misalignment in two images can be corrected. The next layer summarizes these 25 comparison values for each position to a single value. Then, one more convolutional layer is used to calculate relationships across these neighborhood differences. Then, there is a fully-connected layer to obtain the final feature vector to decide on. Finally, a fully-connected layer with two softmax units is used to output the final probability. Modifications that we made on the network architecture are explained

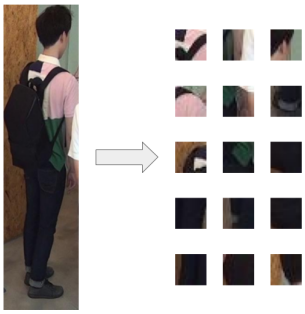


Figure 4.2. Example of an input image cropped into available body parts.

in detail in Section 4.2.

4.1. Extraction of Body Parts

OpenPose [11] is a technique to detect 2D pose information of multiple persons in an image. They use a bottom-up approach, meaning that, first, all potential keypoints in the image and their relationships are estimated. Then, pose information of each person in the image is given as the program output. This way, they achieve to separately detect multiple persons who have overlapping body parts in the image. Also, this allows the program to run in real-time. They publicly released their code.

We applied OpenPose technique to the input images and obtained the detected keypoints. Keypoints detected by OpenPose were shown in Figure 3.4. The surrounding area of each keypoint is defined as a body part. An example of image cropping can be seen in Figure 4.2. These body parts are inputs of individual networks in our person re-id system.

4.2. Network Architecture

The architecture of our network follows from the architecture of Ahmed *et al.* [12]. Since the input of the network is changed from the original image to a part of the image containing a body part, the input size is changed from 160×60 pixels to 20×20 pixels. Also, one of the convolutional layers is removed to reduce the network size. The number of filters, filter sizes in each convolutional layer and the number of nodes in

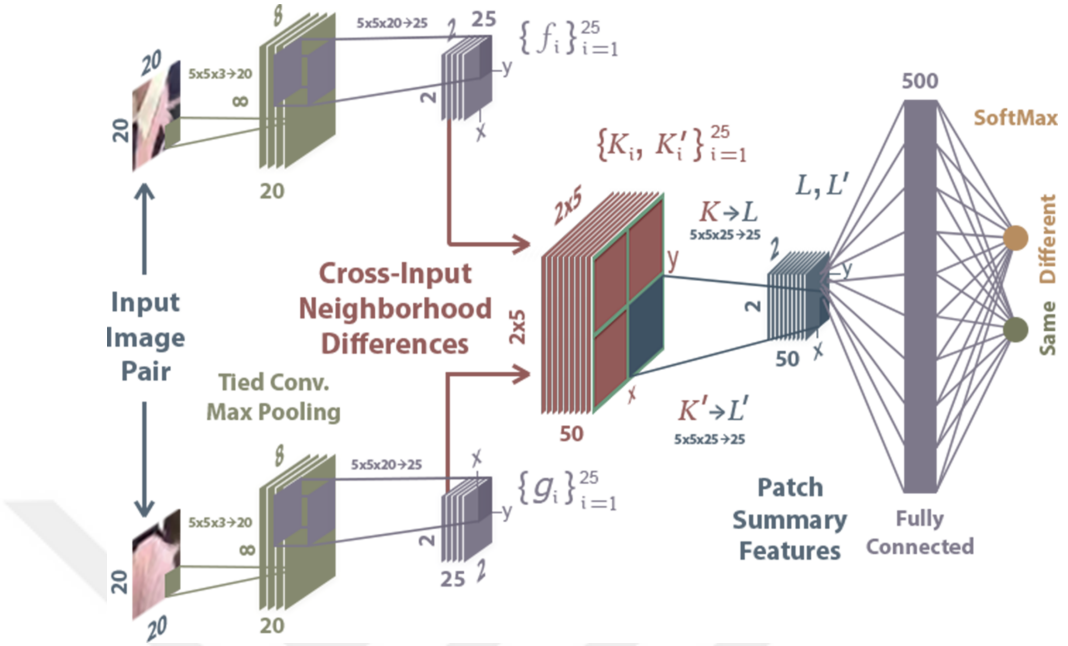


Figure 4.3. Proposed architecture which is a modified version of the architecture of Ahmed *et al.* [12].

fully connected layers are the same as Ahmed *et al.* [12].

The network takes a pair of input images of size 20×20 and they are separately given to two consecutive convolutional layers. Layers which process each image share weights in order to search for similar features in the given input images. Then extracted features are given to the *cross-input neighborhood differences layer* which calculates distances between two feature maps by sliding a 5×5 window in each feature map. For values in each location in one feature map, the difference with values in the window around the same location in the other feature map is calculated. The reason for taking the difference around a neighborhood is that this gives robustness to positional differences of extracted features from two input images. The resulting distances map is given to a convolutional layer. In this convolutional layer, filters are applied with a stride of 5, so that distances to 5×5 window of neighbors are summarized into only one feature. Finally, a fully-connected layer with 500 nodes combines all the information together and gives it to a final fully-connected layer with 2 nodes, which output the probability that the two images belong to the same person. After each convolutional layer, max-pooling is applied. The modified version of the architecture can be seen in Figure 4.3.

4.3. Fusion of Individual Model Results

Openpose [11] outputs the locations of 18 keypoints, therefore we have 18 individual models. Each model outputs the probability of the two input images belonging to the same person, $p(y = 0)$. Three different methods are investigated for fusing these individual model results and best performing fusion method is chosen to be used in the experiments.

4.3.1. Fusion Method 1

The fused probability is calculated by averaging the outputs of individual models:

$$p(y|A, B) = \frac{1}{K} \sum_{k \in \mathcal{I}} p_k(y|A_k, B_k) \tag{4.1}$$

where $y \in \{0, 1\}$, p_k is the output of the kth model, $\mathcal{I}$ is the set of the indexes of available body parts in both of the input images, A and B are the input image pair, A_k and B_k are k^{th} body parts in the image A and B, respectively, K is the number of body parts available in both images and $1 \leq K \leq 18$. A model can be run only when the associating body part is available in both of the input images, therefore (4.1) is chosen to handle the case of missing model outputs.

4.3.2. Fusion Method 2

The fused probability is calculated by processing the outputs for one additional fully-connected layer:

$$p(y|A, B) = \sigma \left(\omega_0 + \sum_{k=1}^{18} \omega_k p_k(y|A_k, B_k) \right) \tag{4.2}$$

where $y \in \{0, 1\}$, A and B are the input image pair, A_k and B_k are k^{th} body parts in the image A and B, respectively, p_k is the output of the kth model and if $k \notin \mathcal{I}$, $p_k(y|A_k, B_k) = 0$ where $\mathcal{I}$ is the set of the indexes of available body parts in both of

the input images. See Figure 4.4.

During the training of the fusion layer, stochastic gradient descent technique is used to update the weights. Each epoch is one iteration over the training data. Learning rate is set to $n = 0.01$. In order to prevent over-fitting, early stopping is applied.

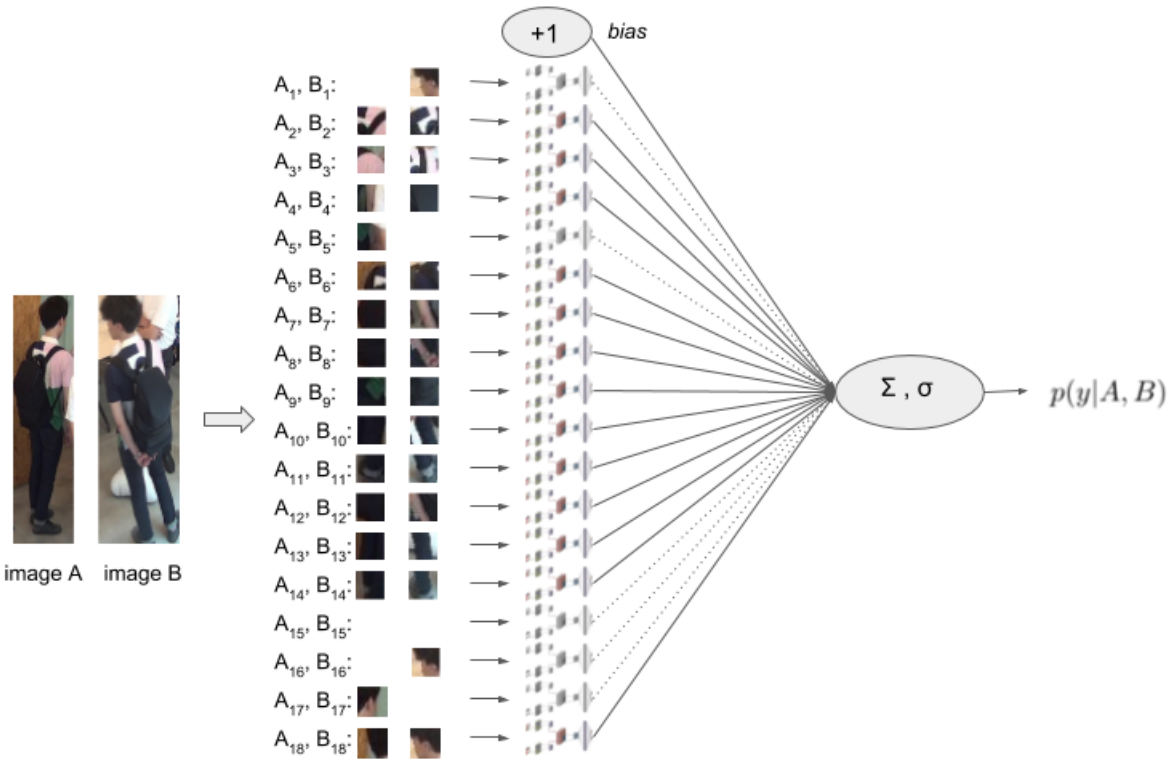


Figure 4.4. Fusion method 2. An additional fully-connected layer is used to fuse the individual model outputs.

4.3.3. Fusion Method 3

The fused probability is calculated by taking a weighted average of individual model outputs according to the keypoint detection confidences associated with each extracted body part:

$$p(y|A,B) = \frac{1}{\sum_{k\in\mathcal{I}} c_{A_k} c_{B_k}} \sum_{k\in\mathcal{I}} c_{A_k} c_{B_k} p_k(y|A_k, B_k) \tag{4.3}$$

where $y \in \{0, 1\}$, A and B are the input image pair, A_k and B_k are k^{th} body parts in the image A and B, respectively, p_k is the output of the kth model, $\mathcal{I}$ is the set of the indexes of available body parts in both of the input images, c_{A_k} and c_{B_k} are the keypoint detection confidence values associated with k^{th} body parts in the image A and B, respectively. See Figure 4.5.

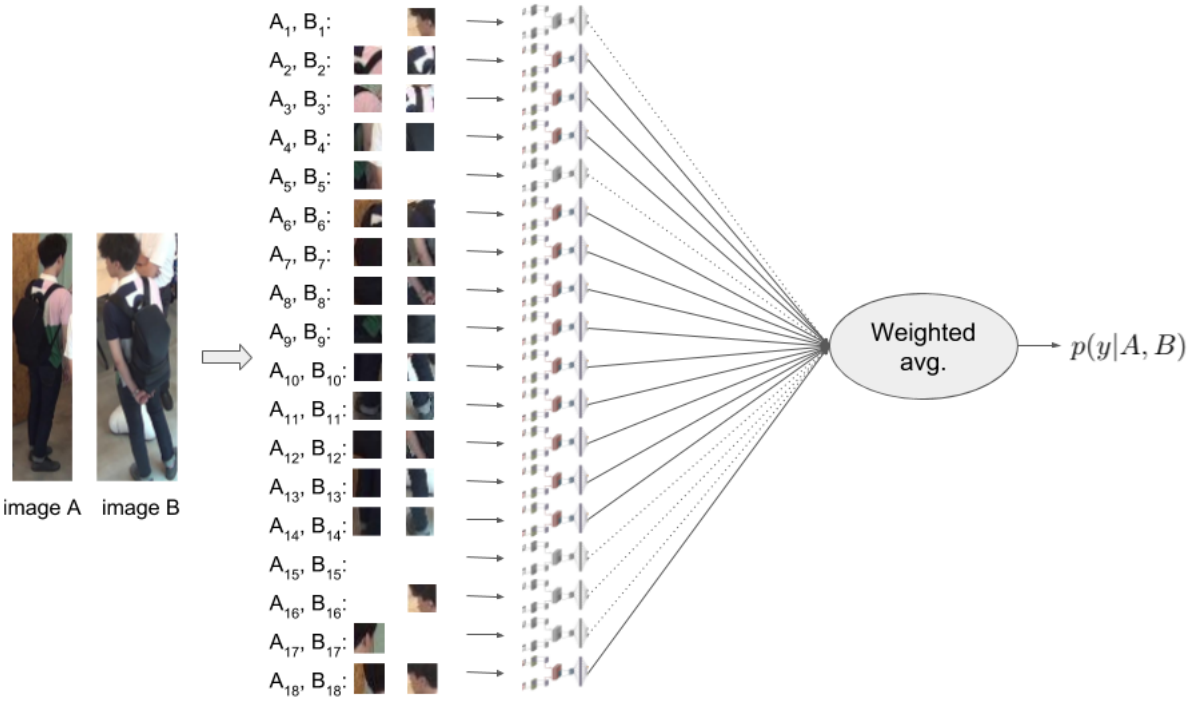


Figure 4.5. Fusion method 3. Individual model outputs are weighted according to the keypoint detection confidence values and then averaged to fuse the individual model outputs.

4.4. Training

The training and validation data are preprocessed with OpenPose [11] and the raw images are then cropped into up to 18 body parts and data belonging to each body part is given to the associated network as training and validation data. Each network has the same architecture but different weights.

Mini-batch stochastic gradient descent [29] technique is used to update the weights. Mini-batch size is set to 150. The learning rate is $n = 0.01$. To prevent overfitting, L2 regularization is applied with weight decay $\lambda = 5 \times 10^{-4}$ and momen-

tum $\mu = 0.9$. Networks are trained for 30,000 iterations.

In addition to L2 regularization, early stopping is also applied. During the network training process, once in 100 iterations training and validation set losses are recorded. Starting from the beginning, networks are trained again until the iteration count which gave the minimum loss on the validation set.

4.5. Fine-tuning

For small datasets, data size is not sufficient to train the networks from scratch. Therefore, networks are trained with a large dataset. Then the weights on the initial layers are kept unchanged and only the last two convolutional layers continued to be trained with a small dataset. We trained the networks for $5,000$ iterations with a smaller learning rate $n = 0.001$.

Early stopping is applied in the fine-tuning stage as well. Starting from the weights learned at the end of the training stage, networks are trained for 5,000 iterations. Once in 100 iterations, the training and validation set losses are recorded. Then, again, starting from the weights learned at the end of the training stage, networks are trained until the iteration count which gave the minimum loss on the validation set.

5. EXPERIMENTS AND RESULTS

Experiments are conducted as single versus single (SvsS) comparisons which means that in the probe and gallery sets, there is only one image belonging to each person. Results are shown in Cumulative Matching Characteristic (CMC) curves. A CMC curve shows the probability of finding the correct match in the first n matches.

5.1. Comparison of Fusion Methods

As explained in the Chapter 4, three different fusion methods are investigated. Only the fusion method 2, which is a fully connected layer with a single node, required a training step. We used CUHK03 data to train the weights (see Figure 5.1).

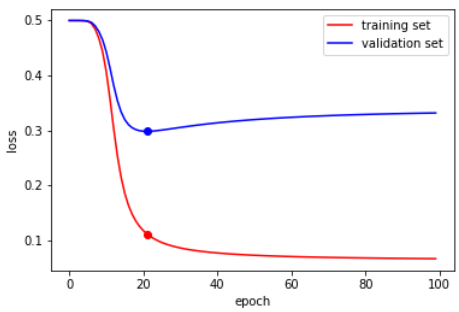


Figure 5.1. Training process of the fusion layer in the fusion method 2.

Fusion methods are compared by their success on CUHK03 data. CMC curves for each method can be seen in Figure 5.2. The fusion method 1 performed the best, therefore it is used in the rest of the experiments.

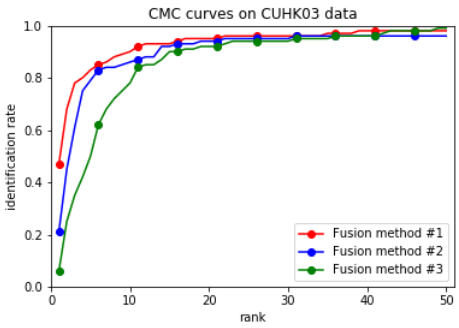


Figure 5.2. Comparison of investigated fusion methods.

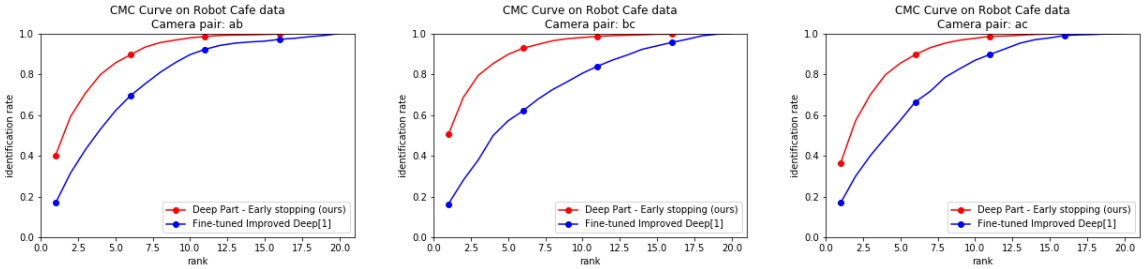


Figure 5.3. CMC curves on Robot Cafe dataset comparing our method to Improved Deep [12] after fine-tuning.

5.2. Experiments on Robot Cafe Dataset

Body part models which are trained with CUHK03 data are fine-tuned with training and validation sets and tested on the test set of Robot Cafe dataset. For each camera pair (a-b, b-c, and a-c), a separate system of networks is created and experiment results are given separately.

In order to make SvsS comparisons, for each person in the test set, a single view from each camera is selected randomly. Then for each pair of cameras, selected data is given to the system and resulted ranks are stored. This process is repeated 50 times.

In order to compare our results, we trained the network proposed by Ahmed *et al.* [12] on CUHK03 and fine-tuned it with Robot Cafe dataset. Results are shown in Figure 5.3. Our method performs much better than [12]. This shows that our method is more robust to challenges such as huge pose changes and missing body parts that Robot Cafe dataset substantially includes.

Fine-tuning processes of each body part model can be seen in Figure 5.4, Figure 5.5 and Figure 5.6 for camera pairs a-b, b-c and a-c, respectively. These graphs show that during the fine-tuning process we were able to improve the body part models related to torso and lower body, whereas the body part models related to face and hands were only slightly improved.

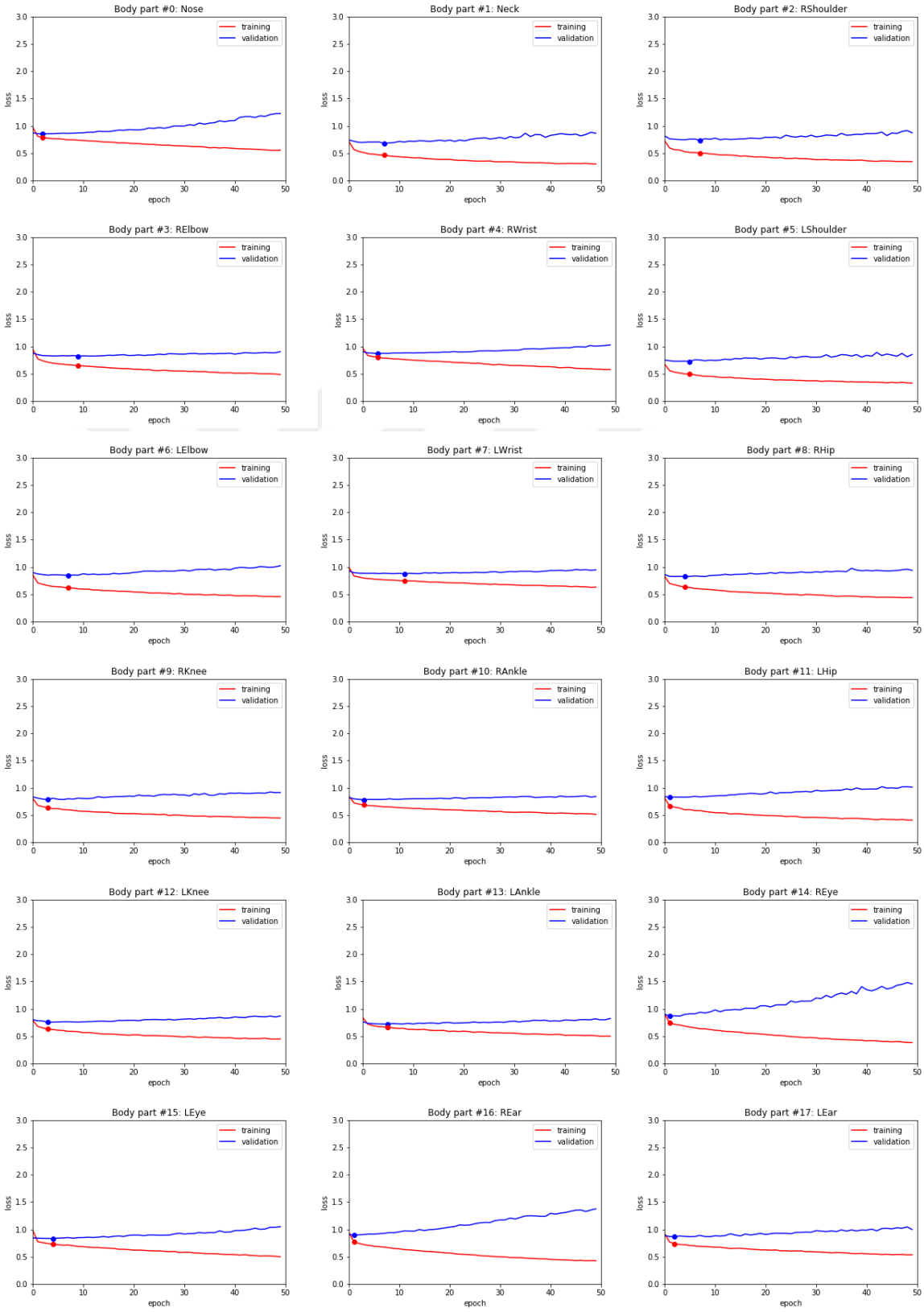


Figure 5.4. Fine-tuning process of each body part model for camera pair a-b.

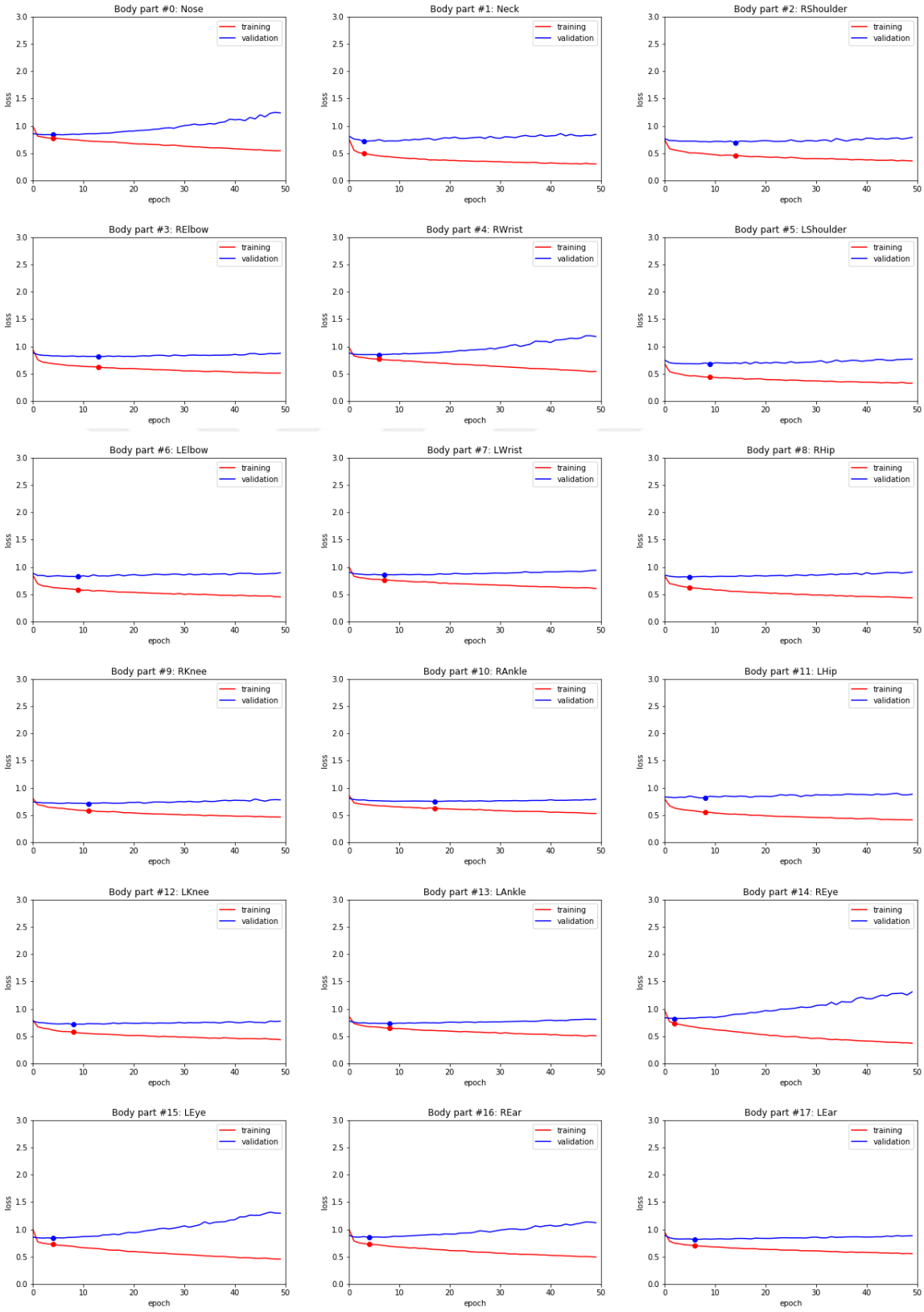


Figure 5.5. Fine-tuning process of each body part model for camera pair b-c.

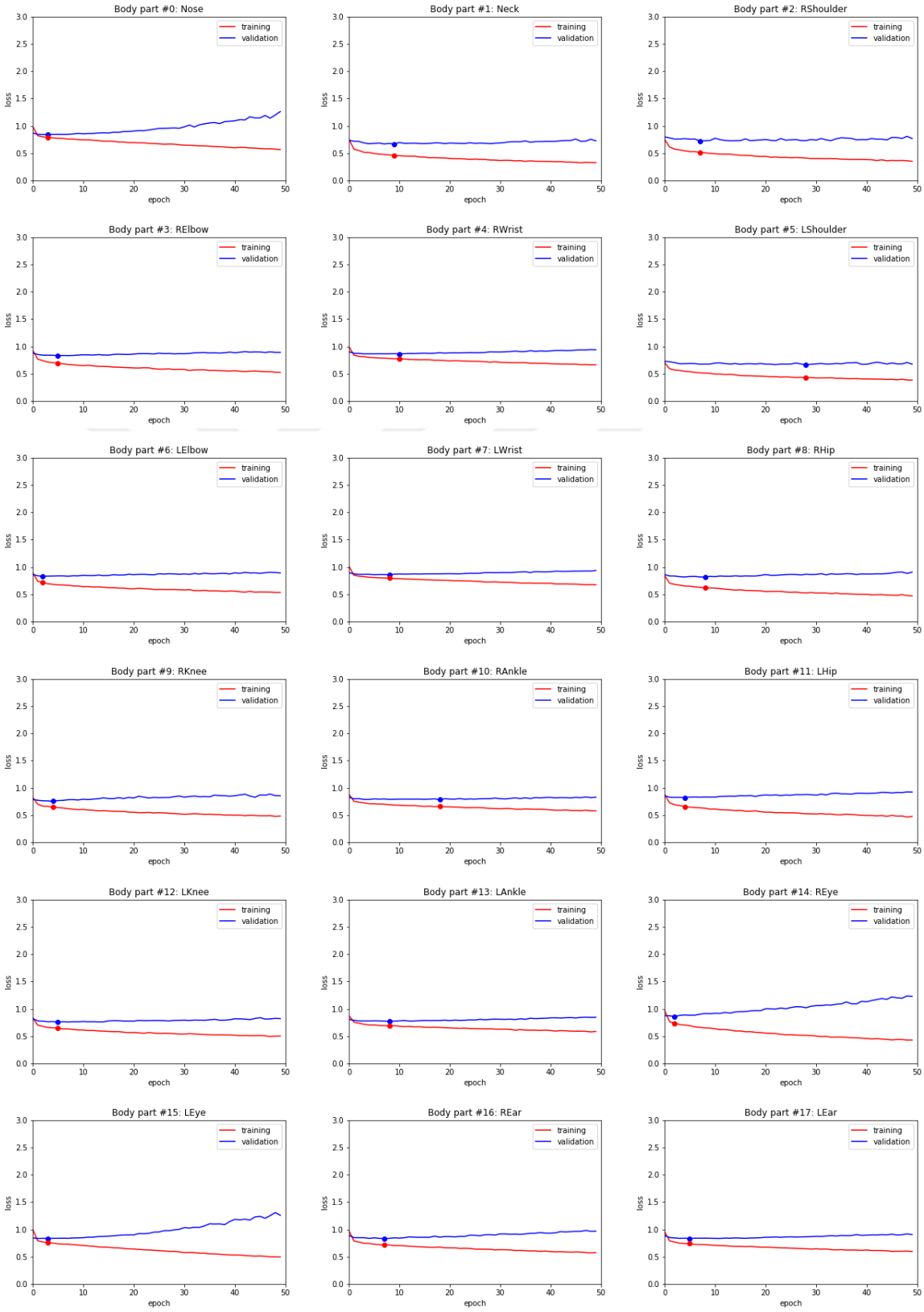


Figure 5.6. Fine-tuning process of each body part model for camera pair a-c.

5.3. Experiments on CUHK03 Dataset

CUHK03 has person images both labeled by hand and automatically detected. In this study, we only used the labeled images. The experiments are conducted on only one data split. Results are shown in Figure 5.7. It is observed that the proposed new model performs better than the baseline and is able to approach [12].

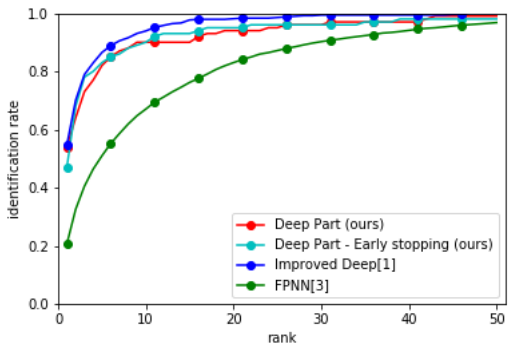


Figure 5.7. CMC curve on CUHK03 dataset comparing our method to some of the previous studies.

Our method could not outperform Improved Deep [12] on CUHK03 dataset. In order to investigate the reasons behind it, we visualized some of the keypoint detection statistics. Figure 5.8 and Figure 5.9 show the detection percentages of keypoints in CUHK03 and Robot Cafe datasets, respectively. It can be seen that detection percentages are more or less the same for both datasets. On the other hand, if we look at the number of keypoints used in comparisons in test sets of CUHK03 and Robot Cafe datasets (see Figure 5.10), we can see that in CUHK03 dataset, less number of keypoints have contributed to the overall output of our method. Since our method is dependent on the keypoint detections for body part extraction, this might have affected our results.

Training processes of each body part model can be seen in Figure 5.11. Poor improvements in the validation set shown in graphs related to the body parts nose, eyes, and wrists suggest that the information on the face and hands does not contribute much to train the related body part models. CMC curves for each body part model can be seen in Figure 5.12. The most successful body part models are neck, shoulders, and

knees, which is expected since these body parts provide the information on upper body and lower body clothing, which gives the most useful cues for discriminating persons from each other.

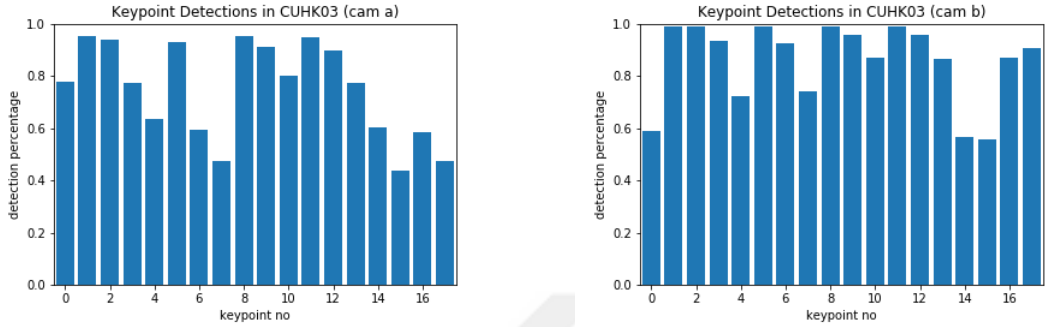


Figure 5.8. Detection percentages of keypoints in CUHK03 dataset.

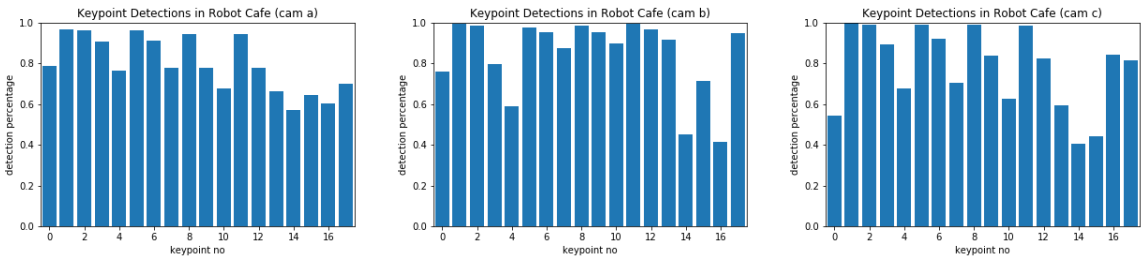
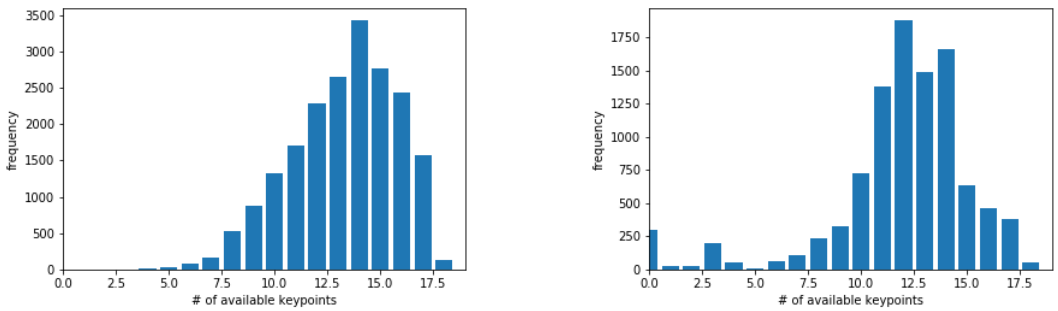


Figure 5.9. Detection percentages of keypoints in Robot Cafe dataset.



(a) Robot Cafe dataset. (b) CUHK03 dataset.

Figure 5.10. Number of body parts used in comparisons.

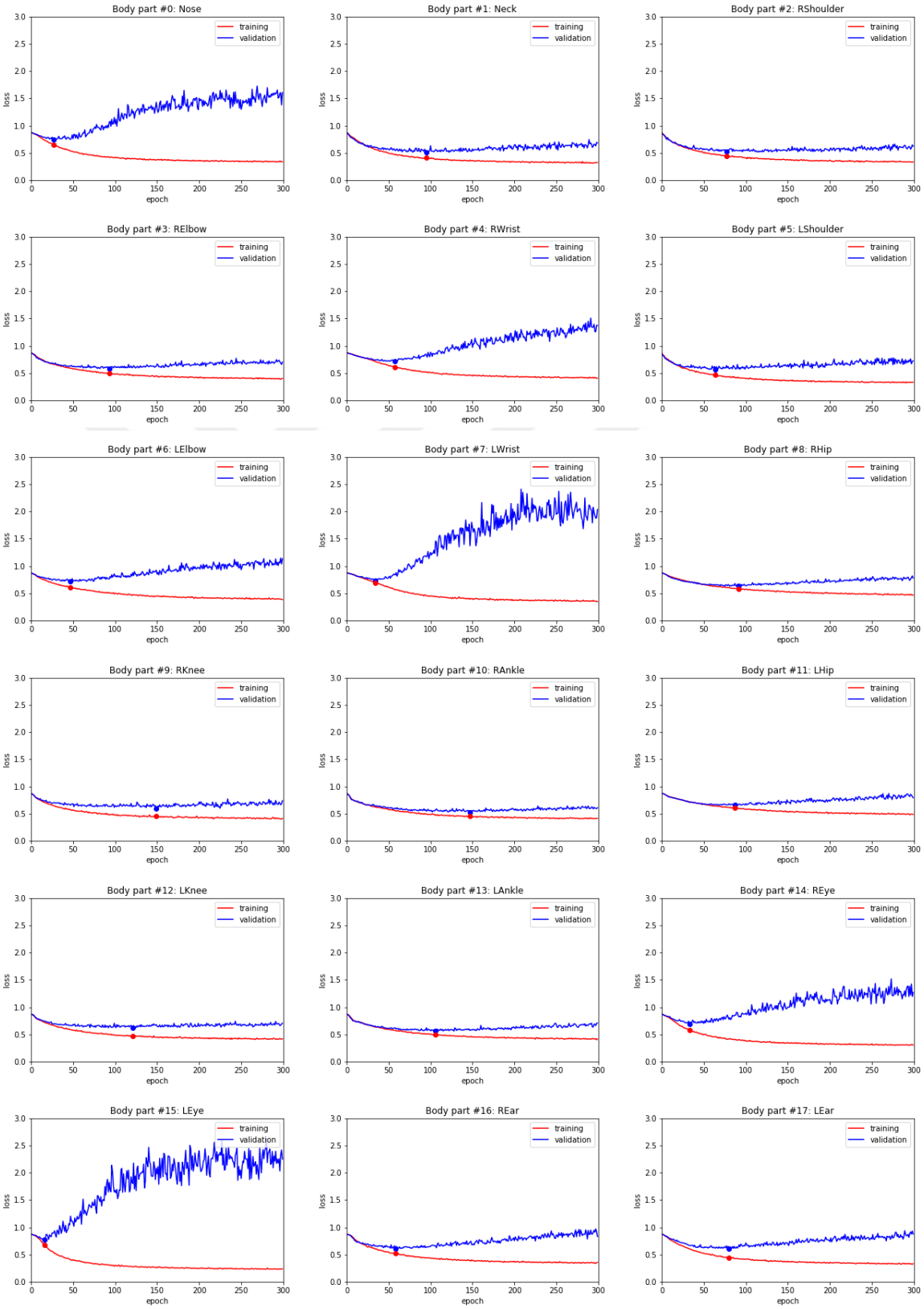


Figure 5.11. Training process of each body part model.

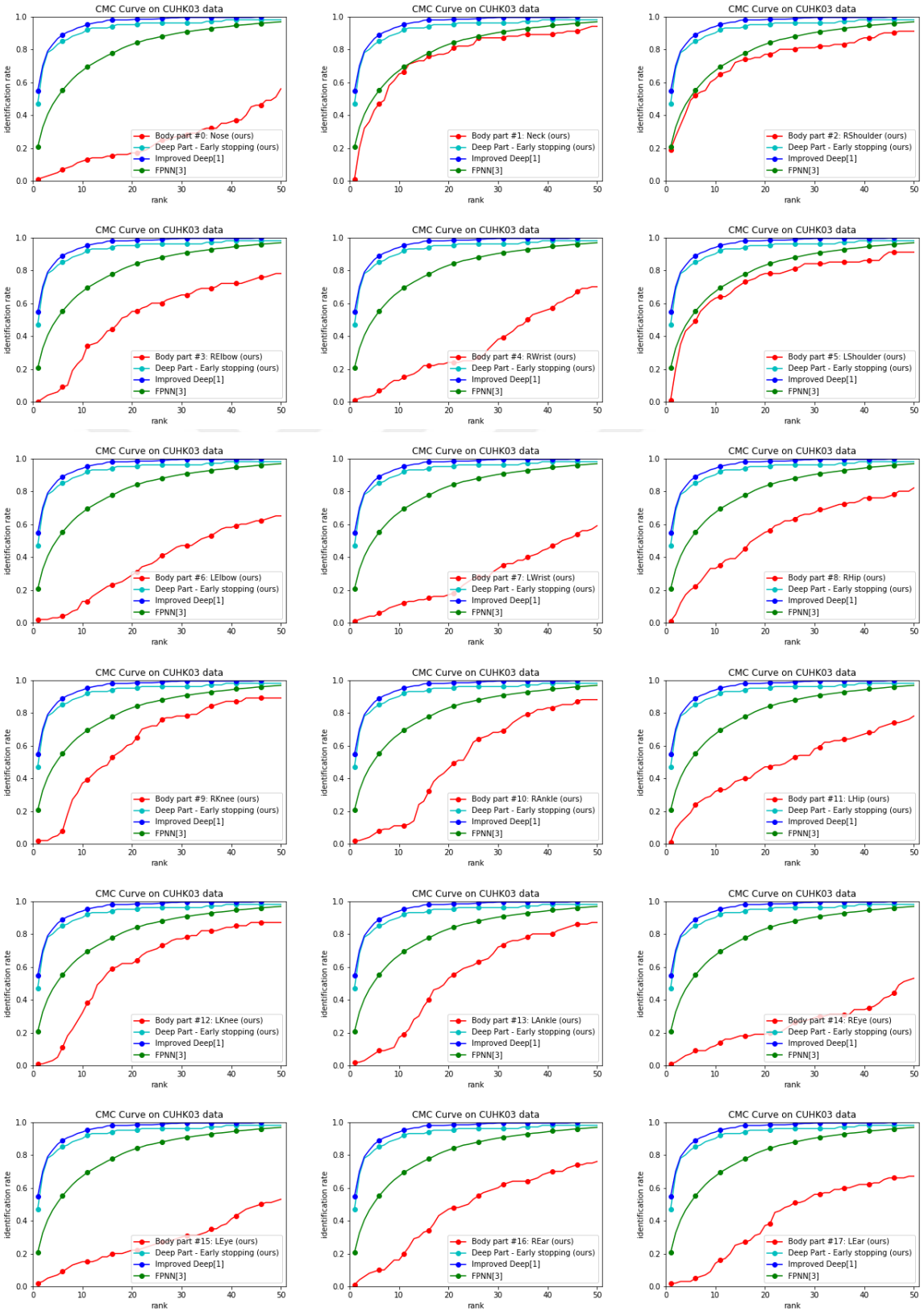


Figure 5.12. CMC curves of individual body part models on CUHK03 dataset.

6. CONCLUSION

In this study, we investigated a part-based method to achieve person re-identification task in order to lift two constraints: a person should be standing and a person's overall body should be available in the image.

Our proposed method starts with the extraction of human body parts. In order to achieve that, we used a human pose estimator which gives the location of up to 18 keypoints together with their confidence values. We trained a deep learning model corresponding to each body part. In order to fuse the results obtained from each body part model, we investigated three different fusion methods. We empirically showed that simply averaging outputs obtained from models associated with the available body parts in the image gave the best results.

In order to evaluate our person re-id system, we collected the Robot Cafe dataset. In Robot Cafe dataset, person images do not necessarily include persons' overall body and persons are recorded while they are standing, sitting or walking, which makes it difficult for person re-id methods which do not take body pose into consideration to perform person re-id accurately.

Our person re-id system was tested on Robot Cafe dataset and the well-known CUHK03 dataset. In order to compare our results on Robot Cafe dataset, we applied the technique of a previous method on the dataset. On CUHK03 dataset, our method performed comparable to some of the previous studies. On Robot Cafe dataset, our method outperformed the compared previous method, which demonstrates that our method is more robust to challenges such as huge pose changes and missing body parts that Robot Cafe dataset substantially includes.

REFERENCES

1. Li, W., R. Zhao, T. Xiao and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159, 2014.

2. Zheng, L., L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable person re-identification: A benchmark", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1116–1124, 2015.

3. Gray, D., S. Brennan and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking", *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, Vol. 3, 2007.

4. Wang, T., S. Gong, X. Zhu and S. Wang, "Person re-identification by video ranking", *European Conference on Computer Vision*, pp. 688–703, Springer, 2014.

5. Schwartz, W. R. and L. S. Davis, "Learning discriminative appearance-based models using partial least squares", *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, pp. 322–329, IEEE, 2009.

6. Albiol, A., J. Oliver and J. Mossi, "Who is who at different cameras: people re-identification using depth cameras", *IET computer vision*, Vol. 6, No. 5, pp. 378–387, 2012.

7. Barbosa, I. B., M. Cristani, A. Del Bue, L. Bazzani and V. Murino, "Re-identification with rgb-d sensors", *European Conference on Computer Vision*, pp. 433–442, Springer, 2012.

8. Satta, R., F. Pala, G. Fumera and F. Roli, "Real-time Appearance-based Person Re-identification Over Multiple KinectTM Cameras.", *VISAPP (2)*, pp. 407–410, 2013.

9. Layne, R., T. M. Hospedales and S. Gong, "Investigating open-world person re-identification using a drone", *European Conference on Computer Vision*, pp. 225–240, Springer, 2014.

10. Ozdemir, H., Y. Ozasa, H. Saito and L. Akarun, "Person re-identification by deep part learning", *24th International Workshop on Frontiers of Computer Vision (IWFCV), Hakodate, Japan*, 2018.

11. Cao, Z., T. Simon, S.-E. Wei and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", *CVPR*, 2017.

12. Ahmed, E., M. Jones and T. K. Marks, "An improved deep learning architecture for person re-identification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3908–3916, 2015.

13. Komlen, D., T. Lombarović, M. Ogrizek-Tomaš, D. Petek and A. Petković, "Multi-camera person tracking system", *MIPRO, 2012 Proceedings of the 35th International Convention*, pp. 1798–1802, IEEE, 2012.

14. Gray, D. and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features", *European conference on computer vision*, pp. 262–275, Springer, 2008.

15. Ess, A., B. Leibe and L. Van Gool, "Depth and appearance for mobile scene analysis", *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

16. Farenzena, M., L. Bazzani, A. Perina, V. Murino and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features", *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2360–2367, IEEE, 2010.

17. Zheng, W.-S., S. Gong and T. Xiang, "Associating Groups of People.", *BMVC*,

Vol. 2, 2009.

18. Wei, L., S. Zhang, H. Yao, W. Gao and Q. Tian, "GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval", *arXiv preprint arXiv:1709.04329*, 2017.

19. Zhang, X., H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang and J. Sun, "AlignedReID: Surpassing Human-Level Performance in Person Re-Identification", *arXiv preprint arXiv:1711.08184*, 2017.

20. Yao, H., S. Zhang, Y. Zhang, J. Li and Q. Tian, "Deep Representation Learning with Part Loss for Person Re-Identification", *arXiv preprint arXiv:1707.00798*, 2017.

21. Zhao, H., M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1077–1085, 2017.

22. Li, W., R. Zhao and X. Wang, "Human reidentification with transferred metric learning", *Asian Conference on Computer Vision*, pp. 31–44, Springer, 2012.

23. Zheng, L., Y. Huang, H. Lu and Y. Yang, "Pose invariant embedding for deep person re-identification", *arXiv preprint arXiv:1701.07732*, 2017.

24. Schmid, C., "Constructing models for content-based image retrieval", *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 2, pp. II–II, IEEE, 2001.

25. Fogel, I. and D. Sagi, "Gabor filters as texture discriminator", *Biological cybernetics*, Vol. 61, No. 2, pp. 103–113, 1989.

26. UK Home Office, *i-LIDS multiple camera tracking scenario definition*, 2008, https://www.gov.uk/guidance/imagery-library-for-intelligent-

detection-systems, accessed at June 2018.

27. Xiao, Q., K. Cao, H. Chen, F. Peng and C. Zhang, "Cross Domain Knowledge Transfer for Person Re-identification", *arXiv preprint arXiv:1611.06026*, 2016.

28. Cheng, D., Y. Gong, S. Zhou, J. Wang and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335–1344, 2016.

29. Bottou, L., "Stochastic gradient tricks", *Neural Networks, Tricks of the Trade, Reloaded*, Vol. 7700, pp. 430–445, 2012.