PERSON RE-IDENTIFICATION &

VIDEO-BASED HEART RATE ESTIMATION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Dahjung Chung

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Edward J. Delp, Chair

School of Electrical and Computer Engineering

Dr. Jan P. Allebach

School of Electrical and Computer Engineering

Dr. Mary L. Comer

School of Electrical and Computer Engineering

Dr. Fengqing M. Zhu

School of Electrical and Computer Engineering

Approved by:

Dr. Dimitrios Peroulis

Head of the School Graduate Program

To my parents

with my deepest gratitude

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my advisor, Professor Edward J. Delp. I am grateful to him for the opportunity of becoming a member of the Video and Image Processing Laboratory (VIPER) laboratory for my Ph.D study. Under his supervision, I learn how to define and solve challenging problems as an individual researcher. I thank him for guiding me throughout my Ph.D study and continuing to push me forward for better achievements.

I am also grateful to my committee members, Professor Jan P. Allebach, Professor Mary L. Comer and Professor Fengqing Maggie Zhu, for their advice to complete my study.

I would like to thank Dr. Khalid Tahboub for being a great teammate for ReID project. I am grateful for his friendship and support during my Ph.D. study. I would like to extend my gratitude to all my former and current VIPER lab members for their friendships and supports: Dr. Albert Parra Pozo, Dr. Neeraj J. Gadgil, Dr. Joonsoo Kim, Dr. Yu Wang, Blanca, Dr. Chichen Fu, Dr. Shaobo Fang, Dr. Javier Ribera Prat, Dr. David Joon Ho, Blanca Delgado, He Li, Chang Liu, Sriram Baireddy, Enyu Cai, Alain Chen, Di Chen, Qingshuang(Cici) Chen, Yuhao Chen, Jeehyun Choe, David Gera, Jiaqi Guo, Shuo Han, Hanxiang (Hans) Hao, Jiangpeng He, Jnos Horvth, Han Hu, Soonam Lee, Runyu Mao, Daniel Mas Montserrat, Ruiting Shao, Zeman Shao, Changye Yang, Sri Kalyan Yarlagadda, Yifan Zhao.

I would also like to thank to Engineering Projects In Community Service (EPICS) program. I was a graduate teaching assistant for EPICS for two years and had a privilege to communicate and help undergraduate students for their research projects and community service through EPICS. This opportunity taught me to be a better engineer by experiencing the design process.

I would like to express my gratitude to my master's degree advisor, Prof. Yoonsik Choe. I appreciate his kindness and supports during my master's program. I learned a lot of things in my research and career from the opportunities he offered. His guidance also help me to be an independent researcher.

I want to show my thankfulness to all of my friends in Purdue and South Korea for their support and encouragement. With the great memories I shared with them, I was able to achieve this journey. I would like to express my special thanks to Kyung-min Hong, for all his love and support.

Last, but not least, I would like to express my sincere appreciation to my parents. They encouraged me to explore new opportunities in life and career and supported me for my whole life. This journey would not have been possible without their endless love and supports.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Chung, Dahjung Ph.D., Purdue University, August 2019. Person Re-Identification & Video-based Heart Rate Estimation. Major Professor: Edward J. Delp.

Estimation of physiological vital signs such as the Heart Rate (HR) has attracted a lot of attention due to the increase interest in health monitoring. The most common HR estimation methods such as Photoplethysmography(PPG) require the physical contact with the subject and limit the movement of the subject. Video-based HR estimation, known as videoplethysmography (VHR), uses image/video processing techniques to estimate remotely the human HR. Even though various VHR methods have been proposed over the past 5 years, there are still challenging problems such as diverse skin tone and motion artifacts. In this thesis we present a VHR method using temporal difference filtering and small variation amplification based on the assumption that HR is the small color variations of skin, i.e. micro blushing. This method is evaluated and compared with the two previous VHR methods. Additionally, we propose the use of spatial pruning for an alternative of skin detection and homomorphic filtering for the motion artifact compensation.

Intelligent video surveillance system is a crucial tool for public safety. One of the goals is to extract meaningful information efficiently from the large volume of surveillance videos. Person re-identification (ReID) is a fundamental task associated with intelligent video surveillance system. For example, ReID can be used to identity the person of interest to help law enforcement when they re-appear in the different cameras at different time. ReID can be formally defined as establishing the correspondence between images of a person taken from different cameras. Even though ReID has been intensively studied over the past years, it is still an active research area due to various challenges such as illumination variations, occlusions, view point

changes and the lack of data. In this thesis we propose a weighted two stream training objective function which combines the Siamese cost of the spatial and temporal streams with the objective of predicting a person's identity. Additionally, we present a camera-aware image-to-image translation method using similarity preserving StarGAN (SP-StarGAN) as the data augmentation for ReID. We evaluate our proposed methods on the publicly available datasets and demonstrate the efficacy of our methods.

1. INTRODUCTION

1.1 Video Heart Rate Estimation

Over the past several decades, physiological vital signs, such as heart rate, respiration rate, have been widely used for the human health monitoring [1]. Especially, the Heart Rate (HR) is an important physiological sign since we can monitor our cardiac activity through it and cardiac diseases are directly connected to the human life.

In order to define the human HR, we introduce how human heart operates. The human heart has left and right sides which operate as a separate pump [2]. The left side of the heart receives oxygenated blood from the lungs and pumps in through arteries to the body tissues. The right side of the heat receives deoxygenated blood from the veins and pumps it to the lungs for oxygenation. A heart cycle (a heart beat) can be defined as a complete blood circulation: the left side pumps to the lung while the right side of the heart pumps blood to the body and lungs. The Heart Rate (HR) is defined as the number of heart beats per unit of time (beat per minute = bpm).

Human HR can vary from 40 bpm to 200 bpm based on many factors such as human body status, age or gender. HR can be divided into two types of range based on the current state of the human body:

1. Basal or Resting heart rate range

   The resting heart rate is defined as the heart rate when a person is not under physical activity or in a statement of rest. The normal range of resting heart rate varies based on the person's age. The normal range is 70 to 130 bpm, 60 to 100 bpm and 40-60 bpm for the children age between 1 year to 10 years, adult and athletes respectively [3].

2. Target heart rate or Training heart rate range.

   The target (or training) heart rate range is defined as the heart rate when a person is under physical activity. In this case, normal range of target heart rate differs based on the person's age, gender and type of physical activity. Although there are many variables to take in account, typical wide range of target hear rate is 60 bpm to 180 $\pm$ 20 bpm [4].

Since HR is a crucial sign of monitoring human health, it has many applications such as patient monitoring, diagnosis and athletes monitoring [5]. This emphasizes the importance of heart rate estimation. Next, we will overview two groups of the previous estimation techniques: conventional and video-based methods.

Among the conventional HR estimation methods, Electrocardiography (ECG) is the most common technique in the medical applications. ECG records electrical signal from the heart activity over the specific time to estimate HR signal. Conventional method of HR estimation from ECG signal is called QRS detection method proposed by Köhler *et al.* [6]. QRS waves correspond to the depolarization of the ventricles of the human heart. Due to this characteristic, it can be used as basis for detecting human heart rate. Although ECG signal has a lot of information related to human health monitoring and we can estimate HR accurately, it is not accessible to common people due to the cost.

Another conventional method is Photoplethysmography (PPG). Detection of the cardiac pulse traveling through the body is referred to as Plethysmography (Plethysmos means increase in Greek) [7]. Photoplethysmography (PPG), first introduced in 1930's [8], uses light reflectance or transmission for simple and low-cost HR monitoring method. PPG is an optical technique that can detect the blood volume changes in the microvascular bed of tissue [9]. It has a near infrared light source and a photodetector to measure the small variation of light intensity in the reflected light from the skin. This small variation is synchronized to the cardiac cycle due to the blood volume changes. PPG technique has been used widely in many clinical applications such as pulse oximeters (Heart Rate, Oxygen Rate), vascular diagnostics and dig-

ital beat-to-beat blood pressure measurement systems [10]. Despite it's simplicity and inexpensiveness, PPG still has some challenges. First, PPG has less accuracy than ECG and its performance heavily depends on the light source and different skin tones. Although there have been intensive studies to reduce the motion artifacts in PPG signal [11–13], it is still sensitive to physical motions of the subject since it requires contact to capture the signal.

Although described methods (ECG and PPG) are able to estimate HR accurately, they still have some critical limitations. First, for the accurate measurement, sensors need to be attached to the subjects or patients' body such as fingers or earlobes. In some cases, these kinds of physical contacts are not possible. For instance, patients who have tactile sensitivities or skin burn problem can not tolerate attaching devices to their skin. Second, sensors have cables to transfer the signal. Theses cables constraint the movement of the subject. For the case of long periods of monitoring such as Intensive Care Unit (ICU), movement constraints cause significant inconvenience. Besides, in real life application, if the subject wants to measure the heart rate during workout, this cable limits the movement of the subject. Thus, it is valuable to elaborate the non-contact means of HR monitoring in both medical and real life applications to overcome these drawbacks.

One approach of non-contact HR monitoring is using thermal imaging [14]. This approach is detecting the small temperature changes in skin region of thermal image which is caused by cardiac activity. Although using thermal imaging gives us more information related to cardiac activity, it requires special equipment which is very expensive. Expensive cost regulates this approach in the narrow field such as medical studies or laboratory level of work.

Another approach of non-contact HR monitoring is video-based HR estimation. Video-based HR estimation, known as videoplethysmography (VHR) is a safe and low-cost technique originated from the PPG technique. VHR can be defined as the methods that use human face videos to estimate HR signal.

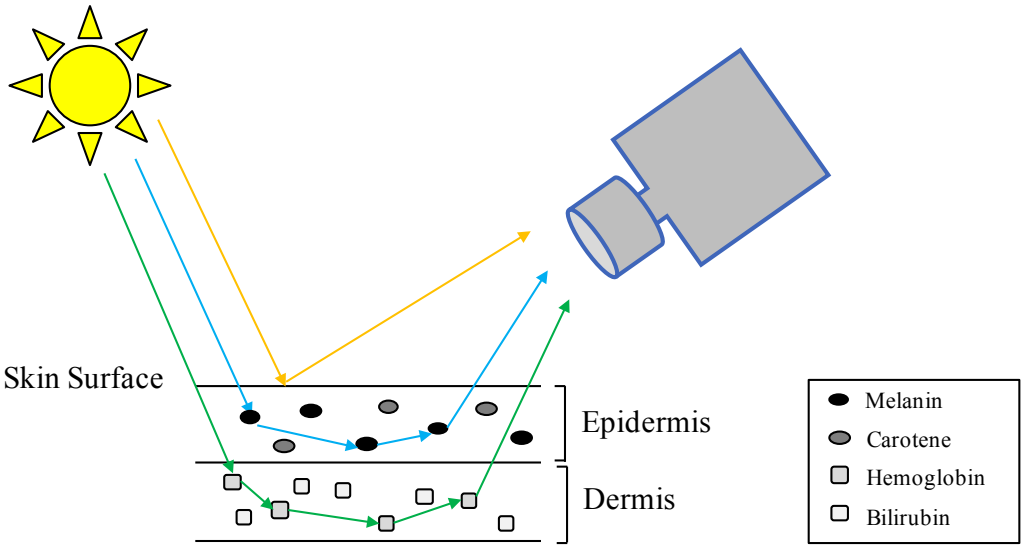


Fig. 1.1.: Basic concept of VHR with human skin optical model [15]

Figure 1.1 shows the basic concept of VHR with the human skin optical model. As shown in Figure 1.1, human skin has two different layers underneath the surface: Epidermis and Dermis. If the light comes to the skin surface, the large amount of the light will be reflected by the skin surface. The rest of the light penetrates into the inside of skin. Some of light is absorbed/reflected in the first layer Epidermis by Melanin or Carotene. The remaining part of the light go deeper to Dermis and be absorbed/reflected by Hemoglobin or Bilirubin. This third light reflection from Hemoglobin in arteries and capillaries reflects the blood volume changes in the vessels which is same as PPG principle. The camera which is the optical device captures the reflected light intensities from the skin to form the image or video sequence. Since one of the reflection carries the heart rate information, we are able to estimate PPG signal from the video by using various processing techniques. Because of its simplicity and low-cost, VHR can be used in many applications including medical patient monitoring or home monitoring using mobile phone camera or webcam. Additionally, it could also inform surveillance systems and provide an alert when someone's heart rate is too high or low.

In past years, Video-based Heart Rate Estimation (VHR) had a lot of attraction and is being intensively studied. Although current VHR methods can estimate HR with certain level of accuracy, there are still challenges to overcome. Current challenges that VHR is facing include:

1. Diverse Skin Tones

   Diverse skin tones make difficulties in VHR methods due to PPG-like principle. Since the basic idea of VHR is the observation of reflected light from the skin, the reflected light from the darker skin tone has very low signal strength and carries less information related to HR.

2. A wide range of typical human HR

   As discussed earlier in this chapter, human HR has a wide range (40 bpm - 200 bpm) based on many variables such as skin tone, age and gender. There might be a large amount of noise in lower or higher frequency band. Noise presence in those bands are challenges in HR estimation.

3. Light Condition Dependency

   Since VHR is originated from PPG, the illumination source is a very important factor. Firstly, ambient light is recommended. Although, we do not have to use ambient light for VHR thanks to previously proposed methods, the low signal strength is still challenging if we have too bright light.

4. Motion Artifact

   Like pulse oximeter (Physical contact PPG device), VHR is also suffering from motion artifact. Although many methods show accurate HR estimations when there is no motion involved, signal-to-noise ratio becomes very low when there is motion involved. This artifact is the most challenging problem in VHR and it still has the room for improvement.

1.2 Person Re-Identification

The security budgets for governments and corporations are increasing over the past decades to protect borders, public transportation, malls, parking lots and homes [16]. As a result, more Closed-Circuit TV (CCTV) systems were employed and the number of video surveillance systems installed has exponentially increased. According to a study by Cisco, Internet video surveillance traffic is projected to increase tenfold between 2015 and 2020 [17].

As the volume of surveillance video has grown exponentially in recent years, making the continues monitoring of surveillance data is impossible. For instance, in airport, we need to have the continuous monitoring and a real-time alert system based on the surveillance video to prevent incidents. However, there will be hundreds or thousands of camera installed in the airport area and it is impossible to have someone watch the feeds from all cameras and detect the suspicious actions or events.

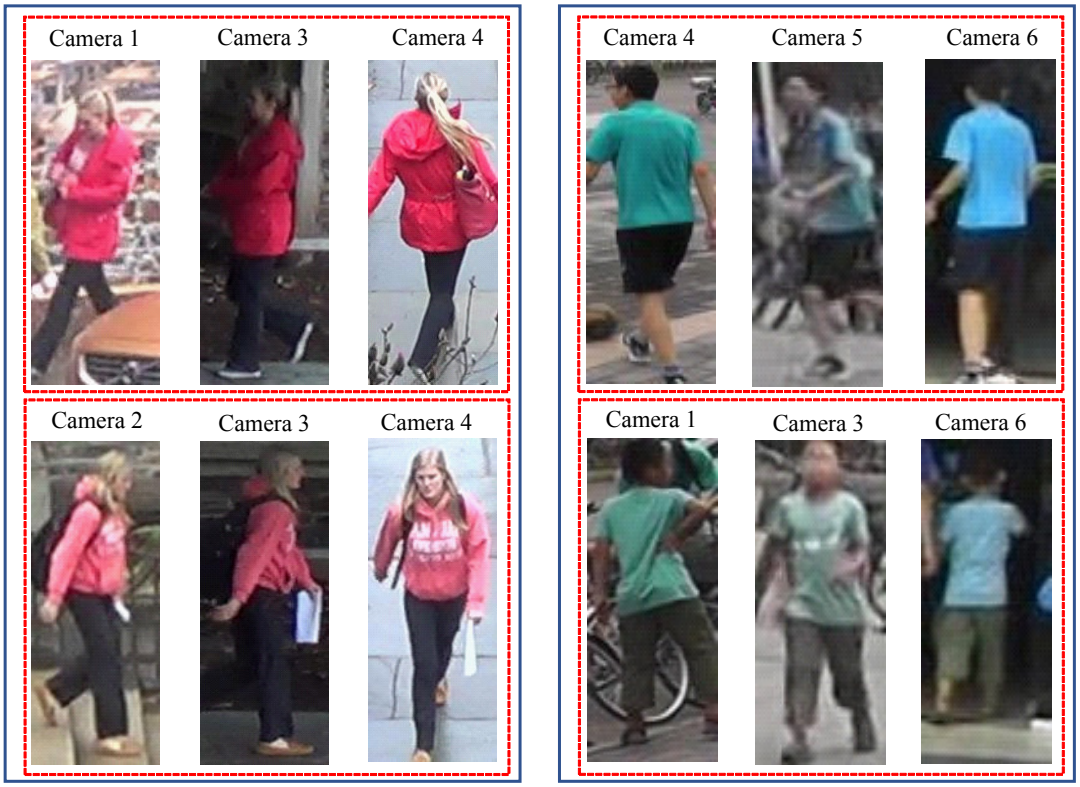
In order to address this issue, intelligent video surveillance system capable of real-time monitoring and automated analysis has been intensively studied in the image processing and computer vision community [18, 19]. The common goal of intelligent video surveillance system is to extract meaningful information efficiently from the large volume of surveillance data. Intelligent video surveillance system has various compute vision tasks to analyze the surveillance video. It includes the object detection, object tracking, action recognition, crowd analysis, human behavior analysis, anomaly detection and person re-identification etc.

Object detection is a task to detect the semantic objects of a certain class such as human or cars in the video [20–22]. Object detection can be used to detect objects to understand the scenes in the videos such as human faces, vehicles. Object tracking is to track the given object as they move around in the frame throughout the time [23, 24]. Object tracking can be applied to track specific person or vehicle in complicated surveillance videos. Action Recognition can be defined as identifying the action occurring in the video and classifying them into action categories [25–28].

Crowd analysis is a method to interpret the movement of groups or objects [29, 30]. This can be used to monitor the crowded scene and to detect the dense crowds. Anomaly detection is a technique to identify the unusual patterns in data that were not expected behaviors [31]. It has many applications such as intrusion detection, fraud detection and public safety monitoring [32].

Lastly, Person re-identification (ReID) is a fundamental task associated with intelligent video surveillance system. ReID can be formally defined as establishing the correspondence between images of a person taken from different cameras at different times [33]. It refers to tracking a person across a network of non-overlapping cameras [34]. Given single/multiple images or a video sequence outlining a person's appearance in the field of view of a camera, ReID is the task of recognizing the same person within the list of images/videos collected from multiple cameras with non-overlapping field of view (gallery). Person re-identification can be used to find the person of interest in the massive volume of surveillance video such as forensic search.

Fig. 1.2.: Sample images of two subjects captured from two different cameras in the PRID2011 [35] and the ILIDS-VID [36] datasets



(a) DukeMTMC-reID [37] (b) Market-1501 [38]

Fig. 1.3.: Sample Images showing challenges related to camera variations in the ReID problem

Even though ReID has been intensively studied over the past years, it is still an active research area due to various challenges. Current challenges in Person Re-identification are :

1. Inter/Intra variations

   Inter-class variation means that different people can look similar across the multiple camera views. Intra-class variation means the same individuals may look very different when they were observed by different camera views. The larger the space and time gap between probe and gallery images, the greater probability will be that people may appear with different clothing or carried objects [39]. In addition, there is a higher chance to have significant illumination

changes which makes hard to distinguish people in the images. Since these variations are multi-modal and complex to model, most of the models trained with a specific pair of cameras do not generalize well to new set of cameras with different viewing conditions [40].

2. Pose variations and Occlusions

   As we have the longer time separation between views, there will be more significant changes in the background. This may also cause a lot of pose variations in the person as well as occlusions by some other objects. Figure 1.2 shows the example of person's pose variations and the occlusions by the other objects

3. Scalability and Different Views

   There will be a lot of different resolutions from different cameras. In addition, each camera has different point of views. These constraints introduces more variations into data and we need to consider these variations when we create a model for ReID.

4. Lack of data and expensive annotation process

   Annotating person identity label along with the camera label is very time consuming. Ideally in ReID, we need to collect images of each person appearing in each camera view at different time in order to learn a robust model to cross-camera and person variations. However, this annotating process gets more time-consuming and extensive as we have a larger camera network.

Figure 1.2 and Figure 1.3 show sample images describing the differences in camera viewpoints and illumination conditions in four different datasets. It demonstrates that it is challenging to distinguish the same person in the images taken from different cameras. For example, in Figure 1.3b, each row 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. This is even challenging for human to distinguish the same person. Most of these challenges

are due to camera variations such as different settings or environments of multiple cameras. In addition, collecting and annotating a large dataset for multiple cameras is a very time-consuming and expensive process.

1.3 Contribution Of The Thesis

In this thesis, we developed new methods for video heart rate estimation and person re-identification. The main contributions of the thesis are listed as follows:

- Improving Video Heart Rate Estimation

    1. We propose a new VHR method using temporal differencing filter and small variation amplification instead of ICA. We observe that green channel signal has all the information we need for HR estimation. Besides, based on PPG principle, HR signal is carried in only small temporal variation. Thus, to compute variation, recursive temporal differencing filter is used. On the temporal differencing filtered signal, we use small variation amplification that amplifies only small variation and reduce large variation.
    2. To get more stable estimations, we propose reduction of signal range for bandpass filter using a cutoff frequency search. Instead of using fixed cutoff frequency for bandpass filter, we search for a tighter and adaptive frequency range starting from the highest peak based on the fact that the highest peak reflects the strongest periodic HR.
    3. We propose an alternative method of skin detection, spatial pruning, for noise removal within face for the future direction. Even though skin detection is performing well on some samples, its performance heavily depends on the lighting condition or the subject's skin tone. Since we compute the average over the skin, if we can not detect any skin points, we will not get any points for future estimation. To get more stable and accurate

information, we propose the use of K-means clustering on the histogram of tracked face.

4. We provide the motion artifact model based on the illumination incident angle principle. We show that this motion artifact exists in the modulation form with HR signal and produces a lot of error in HR estimation. For the future direction, we propose a usage of homomorphic filter for the separation of illumination from a given image to mitigate addressed motion artifact from illumination.

- Person Re-Identification using a Two Stream Siamese CNN

1. We propose a two stream CNN architecture where each stream is a Siamese network. This architecture can learn spatial and temporal information separately. By having two separate networks, each network can learn its own best feature representation.

2. We propose a weighted two stream training objective function which combines the Siamese cost of the spatial and temporal streams with the objective to predict a person's identity. For the ReID task, spatial features are more discriminative than temporal features [41]. The weighted cost function controls the individual contribution of the two streams accordingly. To our best knowledge, this is the first time a weighted two stream cost function is proposed for ReID.

3. We evaluate our proposed method on two publicly available datasets. Our proposed method outperforms or shows comparable results to the existing best perform methods on two public datasets.

- Person Re-Identification using a Similarity Preserving StarGAN (SP-StarGAN)

1. We propose a similarity preserving StarGAN (SP-StarGAN) which is an improvement of StarGAN [42]. To improve the quality of the generated images, we propose to add a identity mapping term and multi-scale structural

similarity (MS-SSIM) to the generator loss. SP-StarGAN can be used not only for ReID data augmentation but also for general multi-domain image-to-image translation. Compare to the previous method (Camstyle [43]), our method has almost 15 times less parameters to train while producing competitive generated image quality as well as competitive accuracy in ReID.

2. For ReID, we use SP-StarGAN to generate more training samples across different cameras. In addition, we propose to employ the Re-Ranking method [44] for ReID as post processing along with SP-StarGAN generated samples in order to improve ReID matching accuracy. We demonstrate that Re-Ranking shows higher performance in ReID accuracy with better quality generated images.

1.4 Publications Resulting From Our Work

1. **D. Chung**, and E. J. Delp, "Camera-Aware Image-to-Image Translation Using Similarity Preserving StarGAN For Person Re-identification," *To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2019, Long Beach, CA.

2. **D. Chung**, K. Tahboub and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," *Proceedings of the International Conference on Computer Vision*, pp. 1983-1991, October 2017, Venice, Italy.

3. **D. Chung**, J. Choe, M. E. OHaire, A.J. Schwichtenberg, and E. J. Delp, "Improving video-based heart rate estimation," *Proceedings of the IS&T International Symposium on Electronic Imaging*, February 2016, San Francisco, CA.

4. J. Choe, **D. Chung**, A. J. Schwichtenberg, and E. J. Delp, "Improving video-based resting heart rate estimation: A comparison of two methods," *Proceedings*

of the IEEE 58th International Midwest Symposium on Circuits and Systems, pp. 1-4, August 2015, Fort Collins, CO.

2. LITERATURE REVIEW

2.1 Video Heart Rate Estimation

In 2008, Verkruysse *et al.* shows that PPG signal can be remotely estimated by using human face video with normal ambient light [7]. In their work, they demonstrate the effect of the spatial averaging over the face to improve Signal Noise Ratio (SNR) of pulse signal. Through their work, the feasibility of VHR has been proven and it has attracted significant attention by many people. Sun *et al.* [45] also demonstrates the feasibility of VHR using the low-cost camera with spatial averaging approach. In this work, Sun shows that we can estimate HR from the video obtained by a high-resolution camcorder or a inexpensive webcam.

Several VHR methods have been developed over the past years based on following two basic assumptions from its PPG-like principle.

1. Small color variations in the cheek/face region reflect blood volume changes (i.e., heart-beats). This is sometimes known as "micro-blushing".

2. Given the rhythmic nature of heart-beats, the color variations will also follow an oscillatory pattern of cardiac cycle.

In the rest of the section, we will review four categories of video-based HR estimation: (1) Independent Component Analysis (ICA) Approach, (2) Motion Detection/Amplification Approach, (3) Chrominance-baed Approach and (4) Other Approaches.

2.1.1 Independent Component Analysis (ICA) Approach

Independent Component Analysis (ICA) approach is the most popular and well-studied approach to estimate HR from the face videos. It is first proposed by Poh

(and Picard) *et al.* [46] in 2010. This method uses Independent Component Analysis (ICA) to decompose 1D signals obtained from the video to separate HR signal. Poh's method begins with face detection to localize the estimation in region of interest (ROI). The average of the pixel intensity of the 60% width of face region is obtained in each RGB frame to form three 1D signals denoted as $x_i(t)$ where $i = 1, 2, 3$. Three 1D signals x_1, x_2 and x_3 are normalized with z-score normalization to form a zero-mean and unit variance signal using Equation 2.1

$$x_i'(t) = \frac{x_i(t) - \mu_i}{\sigma_i} \tag{2.1}$$

where μ_i, σ_i are mean and variance of $x_i(t)$ respectively. Then normalized signals $x_i'(t)$ are decomposed into three underlying source signals using ICA. ICA assumes that the given signals are linear mixture of source signals [47] as shown in Equation 2.2

$$\mathbf{y}(t) = A\mathbf{x}(t) \tag{2.2}$$

where $\mathbf{y}(t) = [y_1(t), y_2(t), y_3(t)]^T$, $\mathbf{x}(t) = [x_1(t), x_2(t), x_3(t)]^T$ and 3x3 mixture matrix A. ICA estimates a decomposition matrix W that maximize the non-Gaussianity of each source. After estimating W, underlying sources $\hat{\mathbf{x}}(t)$ can be determined by Equation 2.3.

$$\hat{\mathbf{x}}(t) = W\mathbf{y}(t) \tag{2.3}$$

Poh's method always choose the second component of ICA output. Then, they compute Power Spectrum Density (PSD) to estimate average HR. The highest peak within [0.75 - 4] Hz is determined as the estimated pulse frequency. However, choosing always second component $\hat{x_2}(t)$ might be inaccurate choice even though they claim that it typically contains strongest periodic signal.

In 2011, Poh (and Picard) *et al.* has extended their own work using ICA approach [48]. Figure 2.1 shows the overall block diagram of Poh's second work. For the remainder of this thesis, this method will be referred as Picard method [48]. Since Picard method will be used as a benchmark for comparison purpose, we are elaborating more details about it.

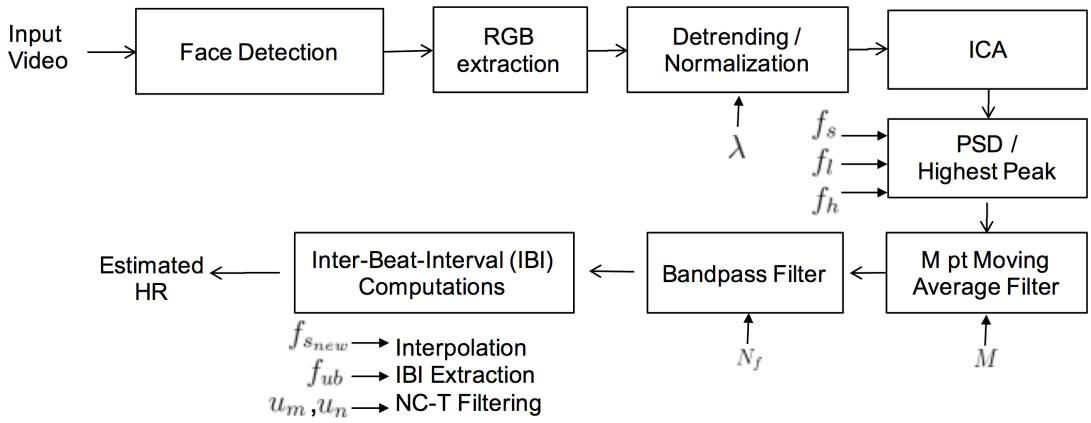


Fig. 2.1.: The block diagram of the Picard method [48].

Figure 2.1 shows the overall flow of the Picard method. Picard starts with the same face detection. The average of pixel intensities within 60% width of detected face is computed to form three 1D signals just like Poh's method. The 1D signals are detrended using a high-pass like filter based on a smoothness priors [49] in order to remove the slow trend in the signal. The parameter for the detrending filter sets the high pass cutoff frequency. Picard denotes this parameter as λ. This parameter is empirically determined by the sampling rate of the data. We set $\lambda = 100$ which corresponds to $0.021 f_s$ where f_s is the sampling rate (in our case, $f_s = 30$ Hz which is the frame rate of our video sequence) for the future experiment. The detrended signal is normalized with same normalization (Equation 2.1) to form a zero-mean and unit variance signal. ICA is used on the three 1D normalized signals to separate the HR signal from the other noise signals using same Equations 2.2 and 2.3. The Power Spectrum Density (PSD) is obtained for the three ICA components. Unlike the Poh's method, Picard chooses the ICA component which has the highest peak in PSD within the resting HR range (0.7 to 2 Hz). This is the choice based on the idea that strongest periodic signal is the our target signal HR among the source signals. After a 5 point moving average filter ($M = 5$), the signal is bandpass filtered using a 128-point Hamming window (filter order $N_f = 127$) with fixed cutoff frequency. The

Low cutoff frequency (f_l) and high cutoff frequency (f_h) are 0.7 and 2 Hz respectively. Next, the bandpass signal is interpolated to a higher sampling frequency $f_{s_{new}} = 256$ Hz using cubic spline interpolation. In the last block of Figure 2.1, the Inter-Beat-Interval (IBI), is the time interval between two peaks in units of seconds, is determined to estimate the HR. The IBI block detects peaks and finds the interval between them to estimate the HR. Finally, the IBI signal is filtered using the NC-T filter [50] in the IBI block with fixed parameters $u_n = 0.4$ and $u_m = 1.0$ Hz. This filter can remove the unstable HR estimation by filtering the rapidly changing values. Poh and Picard methods show the mathematical approach to separate the heart rate signal from the observed noisy signal. However, they have only tested on the face zoomed-in and no motion videos which is not feasible in real life. Besides, ICA has the limitation that we can only recover same number of source signal as the number of input signal.

Monkaresi *et al.* [51] improved Poh's method by choosing the ICA components using K-Nearest Neighbor classification. Monkaresi's method has similar overall flow with Picard method. It begins with face tracking. 60% width of tracked face is used to compute the average in order to form three 1D signals. Three 1D signals go through same detrending, normalization and ICA. Then, noise reduction is done by using comparing the current window estimation with the previous window estimation. Unlike the Picard method, they propose a new estimation method using two machine learning techniques (kNN and linear regression) without component selection. PSD of three ICA components are computed and nine features are obtained from PSD. Six of them include the highest peak frequencies in each component before and after noise reduction. The index of the spectrum peak in the PSD of each component form the rest of features. Using obtained nine features and kNN/linear regression, model is trained and used for estimating HR.

McDuff *et al.* also has extended Picard's method by using using a five band digital camera [52]. In this work, to overcome the ICA limitation described above, they demonstrate that increasing the number of input signal by using a five band camera gives better performance in HR estimation. McDuff uses a five band digital

camera which has Red, Green Blue, Cyan and Orange band sensors. Although this method outperforms their previously proposed method, Picard method, the five band camera is not widely used in practice.

Additionally, Choe *et al.* improves the Picard method [48] by using an adaptive cutoff frequency for the bandpass filter (AFR) to achieve a more stable HR estimate [53]. For the rest of this thesis, this method will be referred as AFR method [53]. In Figure 2.2, the white blocks depict Picard method and the gray blocks illustrate extensions/adaptations used in the AFR method. AFR uses an adaptive cutoff frequency range instead of a fixed cutoff frequency range for the bandpass filter. AFR selects frequencies by targeting those within the typical range of human heart rate and by ignoring oscillatory signals that may be present in the background signal (i.e., lighting, camera vibration).

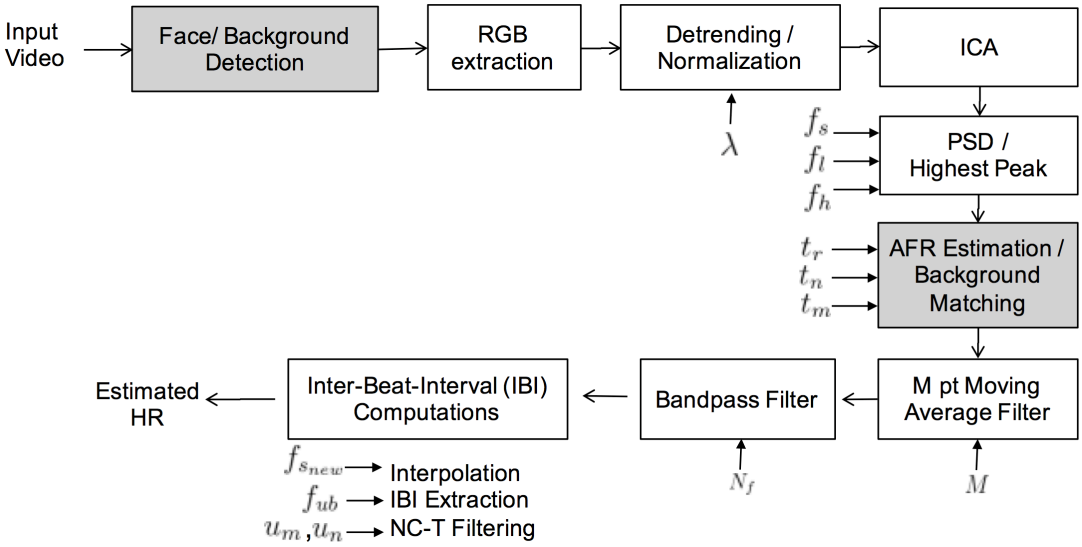


Fig. 2.2.: The block diagram of the our previously published AFR method [53].

AFR begins with face and background region detection. Background region is selected in the flat wall area behind the subject. Two sets of 1D RGB signals from both regions are detrended and normalized. The ICA and PSD processes are the same as the Picard method. Adaptive cutoff frequencies are determined for the bandpass

filter in several steps. First, AFR forms frequency clusters in each PSD from both face and background regions which are the candidates for the cutoff frequencies for bandpass filter. A frequency cluster is a range of neighboring frequencies that are determined by thresholding the PSD. Frequency Clustering, background removal and adaptive frequency range are done in the PSD domain by following steps:

1. Weak signals are ignored when the clusters are formed. If $P[k] < t_r \cdot P_{max}$, then it is determined as weak signal and ignored where $P[k]$ is the PSD and P_{max} is the maximum power. Thresholds $t_r = 0.1(10\%)$ is determined empirically.

2. To form clusters, repeatedly merge the clusters if two clusters are neighbors. t_n [Hz] is used to determine the neighboring clusters. Threshold t_n is chosen as 0.1 Hz (6 bpm) empirically.

3. Compute the sum of the power (energy) of each cluster c_i.

4. Starting from the highest energy cluster in face region, choose a cluster c_{i^*} which meets the criteria: $d > t_m$ where d is given by Equation 2.4

$$d = \sum_{k=0}^{n-1} |P_1[k] - P_2[k]| \tag{2.4}$$

   where P_1 is the PSD of cluster 1 and P_2 is the PSD of cluster 2. Threshold t_m is chosen as 0.4 empirically.

5. Add margins to $f_{i_l^*}$-$f_{i_h^*}$ for the selected cluster c_{i^*} using Equation 2.5 and 2.6

$$f_{a_l} = \max(f_{i_l^*} - t_n, f_l) \tag{2.5}$$

$$f_{a_h} = \min(f_{i_h^*} + t_n, f_h) \tag{2.6}$$

   where i^* is the index of selected cluster, $f_{i_l^*}$-$f_{i_h^*}$ is the frequency range of c_{i^*}. The margin t_n is as same as the parameter from step 2.

Then, ($f_{i_l^*}$-$f_{i_h^*}$) is used as a adaptive frequency range for following bandpass filter block. While Picard is using a fixed frequency range (0.7 - 2 Hz), AFR is using

adaptively narrowed down frequency range for bandpass filter. Once the signal is bandpass filtered, the IBI computation of Picard is done. Through estimating adaptive frequency range with background removal, AFR achieves improvement in terms of accurate estimation.

2.1.2 Motion Detection/Amplification Approach

Rubenstein *et al.* proposed a new method of VHR by detecting subtle changes in the video [54]. The basic idea of his work is that we can amplify the small temporal changes in the videos which are invisible from the human naked eyes.

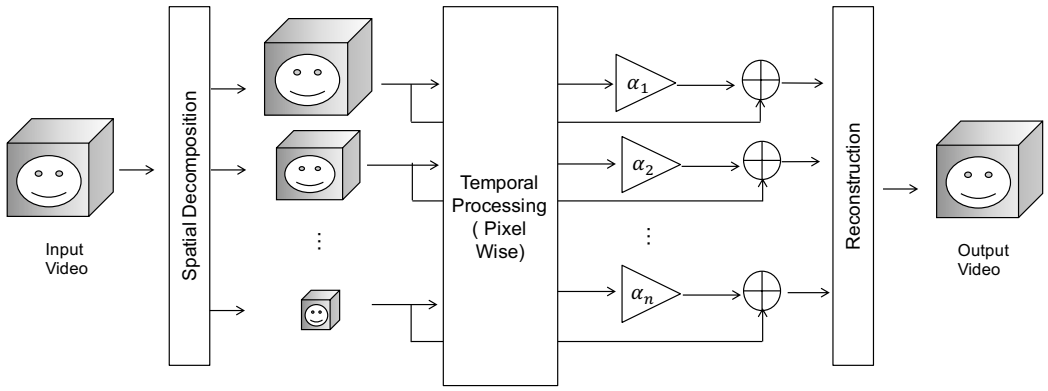


Fig. 2.3.: The block diagram of Rubenstein's method [54]

Figure 2.3 shows the overall block diagram of Rubenstein's method. The sequence of images (frames) are fed into their system as input. In the first block, spatial decomposition, frames of the video are decomposed into the different frequency bands. To decompose the video, they compute a full Laplacian pyramid [55] or Gaussian pyramid by downsampling the frames of video. The purpose of spatial decomposition is to suppress the artifacts and increase signal-to-noise ratio. Then, for the heart rate estimation application, temporal filtering is done using the frequency band of typical HR range 0.4-4 Hz (24-240 bpm). This temporal filtering block produces the small motion magnification effect. They have provided the justification of temporal bandpass filtering by proving the relationship between temporal filtering and motion

magnification. The analysis is given in term of 1D since it can be directly expand to 2D.

$I(x,t)$ denotes the image intensity at position x and time t. As image goes through motion, it can be expressed and approximated with a first-order Taylor expansion,

$$\begin{aligned} I(x,t) &= f(x+\delta(t)) \\ &\approx f(x) + \delta(t)\frac{\partial f(x)}{\partial x} \end{aligned}$$

where $I(x,0) = f(x)$ and $\delta(t)$ is the displacement (subtle motion). Let $B(x,t)$ be the temporal bandpass filtered of $I(x,t)$ given by following equation.

$$B(x,t) = \delta(t)\frac{\partial f(x)}{\partial x} \tag{2.7}$$

Then, motion magnified image will be given

$$\begin{aligned} \tilde{I}(x,t) &= I(x,t) + \alpha B(x,t) \\ &\approx f(x) + (1+\alpha)\delta(t)\frac{\partial f(x)}{\partial x} \end{aligned}$$

By assuming the first-order Taylor expansion holds for the $(1+\alpha)\delta(t)$, we obtain

$$\tilde{I}(x,t) \approx f(x+(1+\alpha)\delta(t)) \tag{2.8}$$

If we magnify the subtle motion $\delta(t)$ directly,

$$\hat{I}(x,t) = f(x+(1+\alpha)\delta(t)) \tag{2.9}$$

we can easily check that $\hat{I}(x,t) = \tilde{I}(x,t)$. So it is justified that subtle motion magnification is same as temporal bandpass filtering by showing the equality of Equation 2.9 to 2.8.

After temporal bandpass filtering, different amplification factor α is computed for each spatial band by using user input parameter λ. Finally, amplified spatial bands are reconstructed into one output video for the final result. If we set our region of interest as face region and use this method, we can see the subtle red color changes which represent the blood perfusion (HR). Although this method is very

intuitive and shows graphical output, it is very sensitive to the subject's motion since it is magnifying the motion. Besides, the performance is heavily based on the input parameters.

Balakrishnan *et al.* [56] estimated HR by detecting subtle head motions. Their basic concept is based on that the cardiac movement of blood from the heart to the head causes the subtle periodic head movement. A person's head video is fed into their system as an input. First, head motion is estimated by feature tracking and the resulting motion model is called trajectories. Then, trajectories are temporal filtered. Next, PCA is used to project the temporal filtered motion information to 1D signals. In PCA, only eigenvectors corresponding to the first five eigenvalues are selected. Finally, to estimate HR, peak detection is used. This work is a novel and creative approach. However, it also turns out that sensitiveness makes it less feasible in real life application.

2.1.3 Chrominance-based Approach

To address the motion artifact problem in ICA and motion detection/amplification based approaches, De Haan *et al.* proposed a chrominance-based VHR method using the ratio of two normalized color signals [57]. Based on the skin optics model shown in Figure 1.1, they have made several assumptions. The reflected light from the skin has two different components.

1. The diffuse or reflection component

   This component is the reflection or diffusion from the layers (Epidermis, Dermis) underneath the skin surface. It carries the color variation information which is synchronized with the cardiac cycle.

2. The specular reflection component (Directly reflected from the skin surface)

   It is the light component reflected by the skin surface which contains the color information of the illumination and no pulse signal.

Besides, they assume that the specular reflection affects all three color channels in the same way (by adding the identical specular fraction to each channel under the white light). Their main assumption is that the non-local intensity variation and specular reflection causes main distortion in the pulse signal. They also assume the standardized skin tone for the subject. They introduce a model for the intensity of a given pixel based on the described assumption [57]

$$C_i = I_{C_i}(\rho C_{dc} + \rho c_i + s_i) \tag{2.10}$$

where the $C = \{R, G, B\}$, I_{C_i} is the intensity of the light source, i is the frame index, ρC_{dc} is the dc component of the reflection, ρc_i is the signal caused by the pulsations and s_i is the specular reflection component. To remove the s_i component which is all same for each color channel (based on their assumption), they propose several different color ratio models such as $XOverY$, Fixed and $X_s min\alpha Y_s$.

For instance, their estimated pulse signal S is modeled with $XOverY$ by Equation 2.11

$$S = \frac{X_n}{Y_n} - 1 \tag{2.11}$$

where $X_n = R - G$ and $Y_n = 0.5R + 0.5G - B$.

De Haan *et al.* has extended his own work for the further improvement [58]. They introduce a new model called 'signature' which is the unique characteristic of the pulse available in the average pixel values of skin region. The normalized blood-volume pulse vector $\vec{P_{bv}}$ is modeled using Equation 2.12

$$\vec{P_{bv}} = \frac{[\sigma(\vec{R_n}), \sigma(\vec{G_n}), \sigma(\vec{B_n})]}{\sqrt{\sigma^2(\vec{R_n}) + \sigma^2(\vec{G_n}) + \sigma^2(\vec{B_n})]}} \tag{2.12}$$

where $\sigma(x)$ is the standard deviation of x.

Wang *et al.* [59] have extended De Haan *et al.* method by exploiting spatial redundancy. Wang's method begins with subject's face tracking for motion compensation denoted as global motion compensation. Within the tracked face, the pixel displacements is estimated using optical flow called local motion compensation. After the motion compensation steps, the difference of same location pixels between adjacent

frames is concatenated to form the signal. Next, spatial pruning is done to select the optical inliers. In spatial pruning block, first skin and non skin classification is done using OC-SVM. Then, in the proposed temporally normalized color space, they get the reliable points by checking the geometric transformation of pixel. Finally, they use adaptive band-pass filtering and PCA decomposition to get the final estimation of HR signal.

2.1.4 Other Approaches

Apart from ICA, many other approaches have been proposed over past years. Another main approach of VHR method is Region of Interest (ROI) based approach. This approach focuses on the locating good region in skin for HR estimation. Tasli *et al.* propose a new ROI based method using adaptive regions within the face while other methods are using fixed ROI [60]. This work begins with face detection and facial landmark feature tracking using the active appearance model [61]. Adaptive ROI is defined based on the facial landmarks and used for computing green channel average to form 1D signal. Next, to obtain the periodic HR signal, they process outlier removal using Equation 2.13 and 2.14

$$p(x, y) \in \text{Outlier, if } \Delta(x, y) \geq 3 * \sigma(R) \tag{2.13}$$

$$\Delta(x, y) = ||p(x, y) - \mu(R)||^2 \tag{2.14}$$

where $p(x, y)$ is the pixel intensity in the selected region R and $\mu(R)$, $\sigma(R)$ are the mean and variance of region R. Then, normalization and detrending are done for final estimated HR signal.

Feng *et al.* propose another ROI based method using dynamic ROI and K-means Clustering [62]. It begins with a fixed ROI region on the skin region. Then, ROI is divided into M x N non-overlapped square blocks. For each block, average of green channel pixel intensity is computed and it is temporal filtered with bandwidth (0.75

- 4) Hz. Then, M x N 1D signals are obtained and Cross Correlation and SNR are computed for each signal using Equation 2.15 and 2.16

$$\gamma = \max_{u}\left(\frac{\sum_n [f(n) - \bar{f}_u][g(n-u) - \bar{g_u}]}{\sqrt{\sum_n [f(n) - \bar{f}_u]^2 [g(n-u) - \bar{g_u}]^2}}\right) \tag{2.15}$$

where $f(n), g(n)$ are two 1D signals, u is a shift parameter, γ is the correlation coefficient and $\bar{f}_u$, $\bar{g_u}$ is the average of $f(n)$ and $g(n)$,

$$SNR = \frac{S(f_{HR})}{\sum_{f=0}^{\frac{f_s}{2}} S(f) - S(f_{HR})} \tag{2.16}$$

where $S(f)$ is the PSD of given signal $f(n)$, $S(f_{HR})$ is the Power at the estimated HR frequency(highest peak from fixed ROI) and f_s is the sample rate. At the same time, they compute the motion masks to avoid the hair or line points in the ROI. From the motion mask, they remove some blocks which involves motion distortion. Using K-means Clustering with two features (γ, SNR), they choose dynamic ROI to remove some blocks which has low signal strength. Finally, by averaging and overlap-adding, they obtain the final estimated HR signal.

Feng *et al.* has extended their own work based on the optical skin model [63]. To discuss the motion artifact, they model the optical signal of remote PPG as

$$I_i(t) = \alpha_i \beta_i (S_0 + \gamma_i S_0 Pulse(t) + R_0) M(t) \tag{2.17}$$

where $i \in \{\text{R,G,B}\}$, S_0 is the average scattered light from ROI, R_0 is the diffuse reflection light from the surface, α_i is the power of ith color light, β_i is the power of diffuse reflection light from ith color light and $M(t)$ is the motion modulation. First, it tracks the ROI using speeded-up robust features (SURFs) points within detected face and KLT method. Next, 1D signals are obtained by averaging and initial bandpass filtered to remove S_0 and R_0 effect in the model. Then, an adaptive color difference is done between green and red channels using Equation 2.18 (this block called Green Red Difference (GRD))

$$GRD(t) = \frac{I_G(t)}{\alpha_G \beta_G} - \frac{I_R(t)}{\alpha_R \beta_R} = (\gamma_G - \gamma_R) S_0 Pulse(t) M(t) \tag{2.18}$$

where $\alpha_G\beta_G$, $\alpha_R\beta_R$ are estimated from the average pixel intensities $I_G(t)$ and $I_R(t)$. After GRD, to remove motion term $M(t)$, an adaptive bandpass filter is used with the initially estimated HR.

Besides the ROI based approach, there is also a weighted signal approach. Yan *et al.* propose another weighted signal approach method [64] using Red, Green and Blue Channels. In their work, they use the forehead area as their ROI and ROI is tracked in each frame. Next, they compute average over ROI to obtain 1D signals for Red, Green and Blue Channels denoted as $x_1(t)$, $x_2(t)$ and $x_3(t)$ respectively. Then, to maximize the signal-noise-ratio, the fixed weighted averaged of three signals is computed based on the Equation 2.19

$$I(t) = \frac{\sum_{i=1}^{3} w_i x_i(t)}{3} \tag{2.19}$$

where $i = 1, 2, 3$, w_i is fixed weighted averaged signal ($w_1 = 0.59$, $w_2 = 0.3$ and $w_3 = 0.11$). w_i is determined based on the fact that green signal has the most information for HR and blue has the least information. Finally, $I(t)$ goes through wavelet transform for de-nosing purpose and bandpass filtered.

Kumar *et al.* propose a weighted signal approach by combining the color signals with adaptive weights [65]. Their work also starts with the tracking facial landmarks from the detected face. The face is divided into the sub regions and these subregions are tracked using KLT method. Then, sub regions are divided into 20 x 20 pixel blocks. Within 20x20 pixel blocks, the average of green channel is computed to form 1D signals. For each 1D signals, the Goodness metric G_i is computed to estimate signal weights using Equation 2.20

$$G_i(PR) = \frac{\int_{PR-b}^{PR+b} \hat{Y}_i(f)df}{\int_{B_2}^{B_1} \hat{Y}_i(f)df - \int_{PR-b}^{PR+b} \hat{Y}_i(f)df} \tag{2.20}$$

where $[PR-b, PR+b]$ denotes a small range around the initial estimated pulse rate PR, $[B_1, B_2]$ is the bandwidth of the bandpass filter, $\hat{Y}_i(f)$ is the PSD of the 1D signal

and i is the sub block index. Finally, the final estimated HR signal $\hat{p}(t)$ is computed using Equation 2.21

$$\hat{p}(t) = \sum_{i=1}^{n} G_i \hat{y}_i(t) \tag{2.21}$$

2.2 Person Re-Identification

We introduced the challenges in person re-identification in Section 1.2. Over the past years, the performance of ReID methods have improved by adopting new features, using metric learning techniques, and the use of semantic attributes and appearance models [41, 66–72].

Most of the traditional approaches proposed for person re-identification uses low-level features in the form of color and texture histograms and exploit metric learning to find a distance function in which distances between images from the same class are minimized and distances between different classes are maximized [68, 69, 73–75]. More recently, deep learning based methods have been proposed to improve the person re-identification accuracy [44, 76–79]. In addition, multiple datasets have been made available for testing, these include: Viewpoint invariant pedestrian recognition dataset (VIPeR) [66], person re-identification (PRID2011) dataset [35] and the iLIDS video re-identification (iLIDS-VID) dataset [36]. VIPeR contains a single image per appearance (single shot), whereas recent datasets, such as PRID2011 and iLIDS-VID, contain multiple images per appearance (multi-shot). Larger-scale ReID datasets have been released more recently: Market-1501 [38] and DukeMTMC-reID [37]. Compare to PRID2011 and iLIDS-VID, Market-1501 and DukeMTMC-reid has more number of persons, images per person and number of cameras.

In the rest of this section, we will review five categories of person re-identification: (1) Feature learning approach, (2) Metric learning approach, (3) Deep learning approach, 4) Generative Adversarial Network Approach and (5) Other approaches.

2.2.1 Feature Learning Approach

In person re-identification, the input is assumed to be bounding boxes outlining persons appearances in two different cameras. Each appearance is represented by a single or multiple bounding boxes. A common approach is to divide a bounding box into a number of horizontal strips and to extract low-level features from each strip.

In [66], an ensemble of local features (ELF) is constructed by using the eight color channels corresponding to the three separate channels of the RGB, YCbCr and HSV color spaces with the exception of the value (V) channel. Thirteen Schmid [80] and six Gabor [81] filters are also used to model texture feature. Sixteen bin histograms are constructed for each of the 19 filter responses and for the eight color channels. The histograms are concatenated to form a high dimensional feature vector for each image. Other approach is the use of the local maximal occurrence feature (LOMO) based on multi-scale Retinex to estimate HSV color histograms used for color features [41]. The scale invariant local ternary pattern (SILTP) descriptor is used to model illumination invariant texture [82]. In [67], a hierarchical Gaussian descriptor (GOG) is proposed and is based on the mean and covariance information of pixel features within patches and region hierarchies. Color and texture features are usually concatenated to form a high dimensional feature vector which is used as an input for learning methods.

In the rest of the subsection, we describe some of the best performing feature extraction methods in more details:

Ensemble of Local Features (ELF)

Gray *et al.* [66] proposed a method that learns simultaneously an ensemble of classifiers and a set of discriminative features. Low level features are consist of color histograms and texture information and they are extracted within local patches. Eight color channels corresponding to the three separate channels of the RGB, YCbCr and HSV colorspace are used along with nineteen texture channels. Schmid [80] and Ga-

bor [81] filters were used to extract texture channels by convolving filters with the luminance channel. Schmid filters are formulated as

$$F(r, \sigma, \tau) = \frac{1}{Z} cos(\frac{2\pi\tau}{\sigma}) exp(\frac{-r^2}{2\sigma^2}) \tag{2.22}$$

and are used to learn to viewpoint and pose invariant features. Z refers to normalization constant and τ, σ are hyper-parameters to be tuned. Gabor filters are used with vertical and horizontal strips. Sixteen bin histograms are formed for all the texture filter responses and for the eight processed color channels. The histograms are concatenated to form a high dimensional feature vector. Then, a feature vector is used the input of metric learning methods.

Local Maximal Occurrence Features (LOMO)

LOMO [41] uses the scale invariant local ternary pattern (SILTP) descriptor [82] using HSV color space. SILTP is an improved operator over LBP(Local Binary Pattern) [83]. Even though LBP has a invariant property under monotonic gray scale transforms, it is inefficient with the nosiy images. SILTP improves the invariance to intensity scale changes by employing a scale invariant local comparison tolerance. In addition, LOMO employs the sliding window and the maximal occurrence among all sub-windows in the same horizontal position. Three-scale pyramic was built and the final feature descriptor was achieved by concatenating all the computed local maximal occurrences.

Hierarchical Gaussian Descriptor (GOG)

Matsukawa *et al.* propose to use a hierarchical model stems from the appearance structure of person images [67]. To achieve the feature representation, GOG employs the part-based model [84]. Images are divided into regions by using horizontal stripes of the image. For each region, local patches are extracted. Each patches are described through a Gaussian distribution of pixel features (patch Gaussians). The mean and co-variance are used to model Gaussian distributions of each patch. The final feature descriptor is form by concatenating four descriptors from different color spaces followed by L2 normalization [85].

2.2.2 Metric Learning Approach

We now describe some of the classification/metric learning methods that have been used in ReID. The keep it "simple and straightforward metric learning" method (KISSME) [68] and cross-view quadratic discriminant analysis (XQDA) [41] are widely used metric learning techniques for ReID. Both approaches belong to the class of Mahalanobis distance functions. KISSME, based on a likelihood ratio test, casts the problem in the space of pairwise differences and assumes a Gaussian structure of the difference space [68]. XQDA extends Bayesian faces and the KISSME approach by learning a subspace reduction matrix and a cross-view metric jointly. A closed-form solution is computed by formatting the problem as a generalized Rayleigh quotient and using eigenvalue decomposition [41]. In [86], the training images are selected and re-weighted based on visual similarities with the query image and its candidate set of images. A weighted maximum margin is trained online and transferred from a generic metric to a candidate specific metric. In [87], a multi shot approach is based on the combination of random projections for dimensionality reduction and random forests for classification. A relative distance comparison model which maximizes the likelihood that a pair of correct match has a smaller distance than that of a wrong match pair along with an ensemble strategy is introduced in [88]. In [89], person re-id is formulated as a block sparse recovery problem and in [90] is formulated as a graph matching problem. In [91], images for a person trajectory are clustered hierarchically to mitigate the problems faced by Fisher Discriminate Analysis (FDA). A viewpoint-invariant descriptor along with sub-image rectification and poses estimation is proposed in [92].

In the rest of the subsection, we describe some of the best performing metric learning techniques in more details:

Keep it Simple and Straightforward Metric Learning (KISSME)
The keep it simple and straightforward metric learning (KISSME) method [68] uses

the class of Mahalanobis distance functions. Given two feature vectors x_i and x_j, the Mahalanobis distance function is formulated as

$$d_M^2(x_i, x_j) = (x_i - x_j)^T \ M \ (x_i - x_j) \tag{2.23}$$

where M is a positive semi-definite matrix. d_M refers to the similarity metric between two person images in person re-identification. The goal of the method using the class of Mahalanobis distance functions is to learn the matrix M. KISSME method test the hypothesis H_0 (dissimilar pair) versus H_1 (dissimilar pair):

$$\delta(x_i, x_j) = log \ \frac{p(x_i, x_j|H_0)}{p(x_i, x_j|H_1)} \tag{2.24}$$

The problem is be re-formulated as

$$\delta(x_{ij}) = log \ \frac{f(x_{ij}|\theta_0)}{f(x_{ij}|\theta_1)} \tag{2.25}$$

in the space of pairwise differences $x_{ij} = x_i - x_j$ with zero-mean assumption.

Assuming a Gaussian structure of the difference space, Equation 2.25 can be rewritten as

$$\delta(x_{ij}) = log\left(\frac{\frac{1}{\sqrt{2\pi|\sum_{y_{ij}=0}|}} exp(-1/2 x_{ij}^T \sum_{y_{ij}=0}^{-1} x_{ij})}{\frac{1}{\sqrt{2\pi|\sum_{y_{ij}=1}|}} exp(-1/2 x_{ij}^T \sum_{y_{ij}=1}^{-1} x_{ij})}\right) \tag{2.26}$$

where

$$\sum\nolimits_{y_{ij}=1} = \sum_{y_{ij}=1} (x_i - x_j)(x_i - x_j)^T \tag{2.27}$$

$$\sum\nolimits_{y_{ij}=0} = \sum_{y_{ij}=0} (x_i - x_j)(x_i - x_j)^T \tag{2.28}$$

and x_{ij} is the symmetric pairwise differences.

The likelihood ratio can be reformulated by taking log as

$$\delta(x_{ij}) = x_{ij}^T \Big(\sum\nolimits_{y_{ij}=1}^{-1} - \sum\nolimits_{y_{ij}=0}^{-1} \Big) x_{ij}. \tag{2.29}$$

The matrix M is obtained by re-projection of $\hat{M} = \Big(\sum_{y_{ij}=1}^{-1} - \sum_{y_{ij}=0}^{-1} \Big)$ onto the cone of positive semi-definite matrices.

Cross-View Quadratic Discriminant Analysis (XQDA)

Liao *et al.* [41] extend the Bayesian face and KISSME [68] methods to cross-view metric learning method. The XQDA considers to learn a subspace $W = (w_1, w_2, ..., w_r) \in \mathbb{R}$ with cross-view data while simultaneously learning a distance function. The distance function with the subspace W is defined as

$$d_W(x, z) = (x - z)^T \; W \; (\sum\nolimits_I^{'-1} - \sum\nolimits_E^{'-1}) W^T (x - z) \tag{2.30}$$

where $\sum_I^{'} = W^T \sum_I W$ and $\sum_E^{'} = W^T \sum_E W$. To optimize the projection direction W such that $\sigma_E(w)/\sigma_I(w)$ is maximized, Rayleigh Quotient

$$J(w) = \frac{W^T \sum_E W}{W^T \sum_I W} \tag{2.31}$$

is used. The maximization of $J(w)$ is solved employing the generalized eigenvalue decomposition in a similar way as in LDA.

2.2.3 Deep Learning Approach

Deep learning is a method that uses neural networks consisting of multiple layers and non-linear activation functions. Recently, deep learning [93, 94] has shown significant performance improvement in many different fields such as natural language processing [95, 96], and graphical modeling [95, 97] and computer vision tasks [98]. LeNet [99] was the first successful case of convolutions neural networks developed by Yann LeCun. LeNet was trained on MNIST dataset to perform the digit recognition. Even though LeNet successfully demonstrates the feasibility of Convolutional Neural Network (CNN), it did not get big attention because it was not feasible to be applied to the real-world applications due to the lack of computation power. With the recent progress in graphics processing units (GPUs), deep learning has gotten a lot of attention especially in computer vision community [100]. Besides, many public datasets are introduced such as ImageNet [101, 102], PASCAL VOC [103, 104] for image classification and SPORTS-1M [105] for action recognition. Along with the larger datasets, CNN had shown a significant improvement in performance due to

the ability to learn more sophisticated features. The first few convolution layers in CNN are able to capture lower level features such as edges whereas the deeper convolution layers are able to learn higher level features by incorporating previous layer's features [106, 107]. The convolution layer can be defined as

$$x_{ij}^{l} = \sum_{a=0}^{m_1-1} \sum_{b=0}^{m_2-1} \omega_{ab} y_{(i+a)(j+b)}^{l-1} \tag{2.32}$$

where m_1xm_2 is the filter size, l is the layer number, y is the input to the l-th convolution layer (the output of the previous layer) and ω is the weights to learn. AlexNet [98] was proposed by using larger convolution filters, max pooling and Rectified Linear Unit (RELU) as non-linear activation. RELU is defined as

$$f(x) = max(x, 0) \tag{2.33}$$

where x is the input to the RELU. The max-pooling is the down-sampling layer for reducing the spatial size of the activation maps. In addition, AlextNet employs the heavy data augmentation which is a method to create extra training images by image rotation, translations or horizontal/vertical flipping to prevent over-fitting. VGGNet [108] uses a series of small size convolution filters such as 3 x 3 to reduce the total number of parameters. GoogleNet [109] uses the inception modules to have various receptive field using 1 x 1 convolution filters, 3 x 3 convolution filters, and 5 x 5 convolution filters. More recently, Residual Network (ResNet) is proposed in [110]. He *et al.* [110] propose the residual block which contains two convolution layers with a shortcut connection. The output of the residual block becomes $F(x) + x$ where x is an input and $F(x)$ is the output of convolution layers. ResNet has shown that we can train deeper layers without performance degradation with the shortcut connections.

To train the neural network, many optimization techniques are introduced such as stochastic gradient descent [111] (SGD) with momentum or the Adam optimizer [112]. In addition, to implement the neural network algorithm, many deep learning frameworks such as Torch [113], TensorFlow [114], Pytorch [115], Caffe and Caffe2 [116] are introduced.

Until recently, CNN architectures [94] have not been used for the ReID due to the small size of public datasets. With the release of larger datasets, recent methods have demonstrated the feasibility of the use of CNNs for ReID [117, 118]. A filter pairing neural network (FPNN) is proposed as a unified solution to extract features and learn photometric and geometeric transforms in [117]. In [118], feature extraction layers are followed by a cross-input neighborhood difference layer to compute the differences in feature values across the camera views.

After this initial work, deep learning ReID methods extended [117, 118] and incorporate metric learning and part-based learning [119–121]. In [119], a cosine layer connects two sub-networks and jointly learn color, texture and a similarity metric. Later, Cheng *et al.* [120] employed a multi channel part-based CNN to jointly learn both global and local features of the human body. The network is trained using triplet images and a triplet loss function is used to learn the network model. In [122], single image and cross-image representations are combined in a single network. A deep learning network for learning features from multiple domains is proposed in [121]. A domain guided dropout (DGD) method [121] is shown to improve feature learning. In [123], a Siamese Long Short-Term Memory (LSTM) model that can process image parts sequentially is described. The use of LSTM enables the capability of memorizing the spatial dependency and selectively propagating the context information throughout the network.

More recently, deep learning based methods have been described such as RNN-ReID [124], SVDNet [78], IDE [125], Re-Ranking [44] and TJ-AIDL [79]. To exploit the temporal information for ReID, the use of recurrent neural network (RNN) with the Siamese structure is proposed in [124]. An optical flow image is concatenated to the YUV image and comprises the input to the deep learning network. For the remainder of this thesis, we will refer to the ReID technique proposed in [124] as the RNN-ReID technique. Si *et al.* [126] propose the dual attention match network to learn context aware features while performing the attentive comparison simultaneously. Besides, it uses a Siamese network with triple loss, de-correlation loss and cross-entroy

loss. In [127], deep Siamese attention network is proposed for video-based person re-identification. By employing attention mechanism, the model can learn which parts in which frames are discriminative and important for person-identification. Sun *et al.* [78] propose to use Singular Vector Decomposition (SVD) [128] to optimize the deep learning model.

In addition, Zheng *et al.* [125] proposed ID-Discriminative Embedding (IDE) using ResNet [110] to train a ReID model as a classifier. In [44], they propose a re-ranking method using k-reciprocal Encoding inspired by [129]. This method can be used with any initial ranking. We refer this method [44] as Re-Rank for the rest of this thesis.

In the rest of the subsection, we describe some of the previous deep learning methods in more details:

ID-Discriminative Embedding (IDE) [125]

IDE uses the softmax loss to learn the ID-Discriminative embedding for person re-identification. This method trains a person re-identification model as image classification task. It uses ResNet-50 [110] as the backbone and fine-tunes from ImageNet pre-trained model. ResNet-50 has 34-layers with 3-layer bottleneck block, resulting in a 50-layers in total. Each person ID is used as the class label for the softmax loss.

RNN-ReID [124]

Figure 2.4 shows the overall flow of RNN-ReId method. McLaughlin *et al.* propose to use both YUV image and optical flow images as the input to exploit spatial and temporal information. First, RNN-ReID process each input for different time stamps with CNN. The CNN is consist of three convolution layers and one fully-connected layer. Then the result of CNN is fed into RNN layer for each time stamps. Temporal average pooling is used to summarize the video features into a single vector. For the final training loss, Siamese cost is computed between feature vectors captured from camera A and camera B.

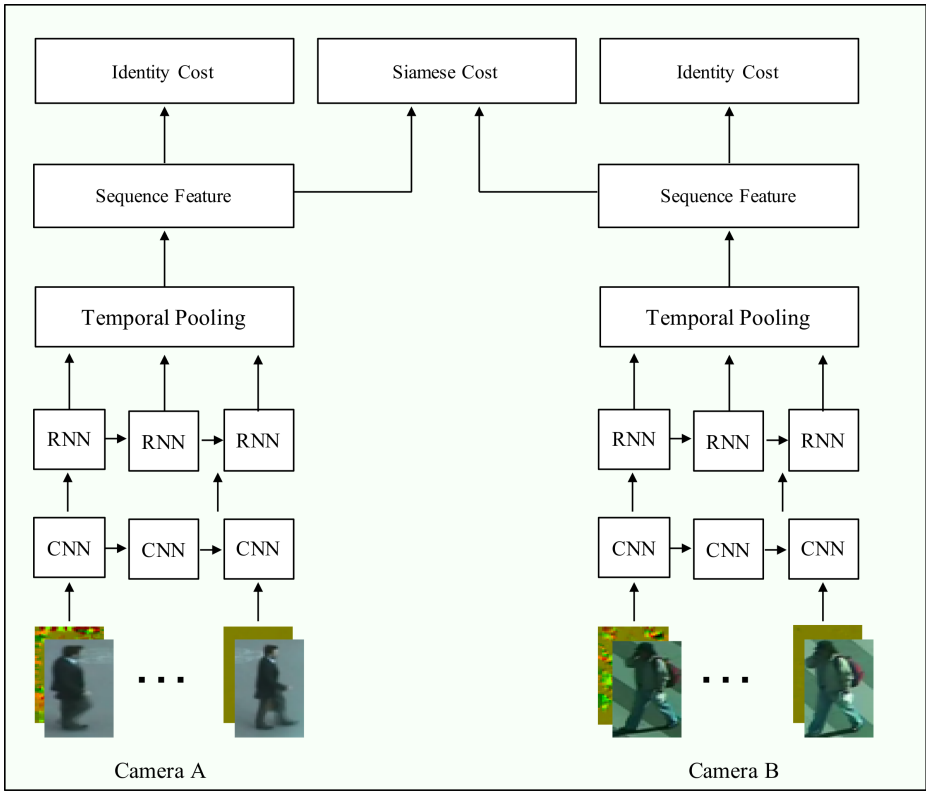


Fig. 2.4.: A overall block diagram for RNN-ReID method [124]

2.2.4 Generative Adversarial Network Approach

Generative Adversarial Networks.

Goodfellow *et al.* [130] proposed the Generative Adversarial Networks (GANs) which learns generative models through an adversarial process that is training a generative model and a discriminator model simultaneously. As Figure 2.5 shows, GANs are consist of two sub-networks: a generator network G and a discriminator network D. The goal of GANs is learning to generate realistic fake images with the generator network while learning to distinguish between real and fake images with the discriminator network. More specifically, G learns the mapping from noise variable z to image space $G(z;\theta_G)$ where θ_G is the generator parameters, whereas D tries to maximize the

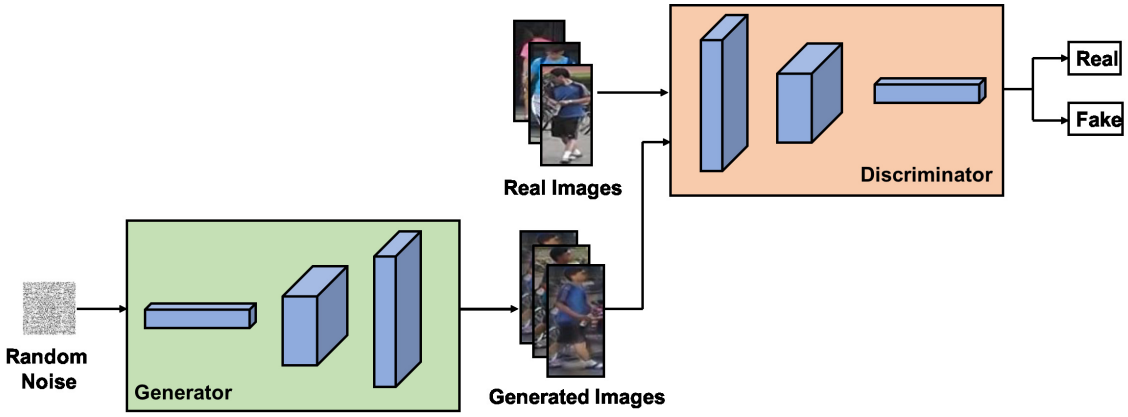


Fig. 2.5.: A overall block diagram of GAN [130]

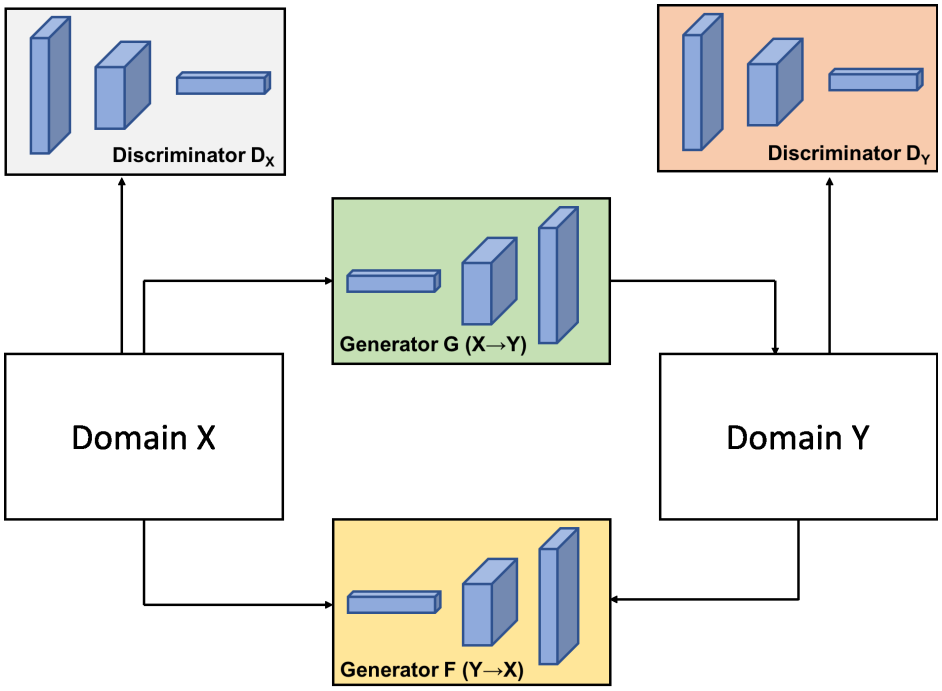
probability of assigning the correct label to both real images and generated images . Therefore, GANs final loss function can be formulated as [130]:

$$\min_{G}\max_{D} V(G,D) = \mathbb{E}_x[logD(x)] + \mathbb{E}_x[1 - logD(G(z))]. \tag{2.34}$$

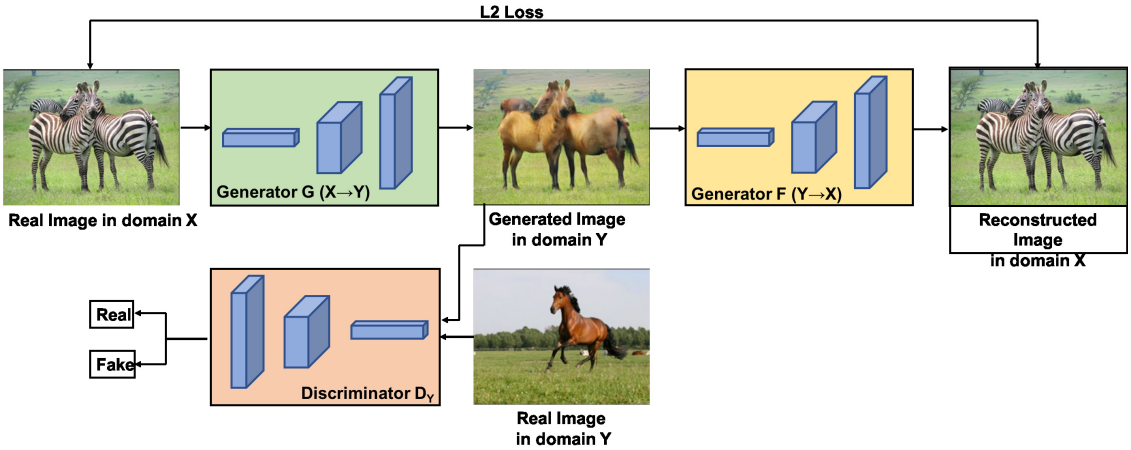
As G and D play the two-player minmax game, G can generate more realistic images while D can distinguish generated images.

In recent years, GANs has been used in many applications including image generation [131] and image-to-image translation [42, 132, 133]. Radford *et al.* [131] introduced deep convolutional generative adversarial networks (DCGANs) that have some architectural constraint for stable training and they have demonstrated the applicability for image generation. One of the extensions of GANs, Pix2Pix [132], used a conditional GANs to learn the relationship between the output and input image for image-to-image translation. Pix2Pix [132] has the limitation that it requires paired training data. To overcome this limitation, a coupled generative adversarial network (CoGAN) was introduced to learn the joint distribution across domains without having the paired training data.

Next, cycle consistency adversarial networks (CycleGAN) [133] employed a cycle-consistency term in the adversarial loss for image-to-image translation without having paired samples. Figure 2.6 shows the overall block diagram of CycleGAN. CycleGAN



(a) A overall flow



(b) Detailed mapping flow

Fig. 2.6.: A overall block diagram of CycleGAN [133]

has two Generators G and F and two discriminators D_X and D_Y. Generator G learns the mapping from domain X to Y whereas generator G learns the mapping from domain Y to X. Discriminators D_X and D_Y learn to distinguish the fake images for each generator. AS shown in Figure 2.6b, the real image in domain X goes through

generator G. We then have generated image in domain Y and this generated image goes through generator F to reconstruct the images in domain X. To preserve a cycle-consistency, L2 Loss is computed between the reconstructed image and the real image. The full objective function of CycleGAN can be formulated as

$$\begin{aligned} L(G, F, D_X, D_Y) = L_{adv}(G, D_Y, X, Y) + L_{adv}(F, D_X, Y, X) \\ + \lambda L_{cyc}(G, F), \end{aligned} \tag{2.35}$$

where where λ controls the relative importance of the two objectives,

$$L_{adv}(G, D_Y, X, Y) = \mathbb{E}_y[log(D_Y(y)] + \mathbb{E}_x[log(1 - D_Y(G(x))], \tag{2.36}$$

$$L_{adv}(F, D_X, Y, X) = \mathbb{E}_x[log(D_X(x)] + \mathbb{E}_y[log(1 - D_X(F(y))], \tag{2.37}$$

$$L_{cyc}(G, F) = \mathbb{E}_x[||F(G(x)) - x||_1] + \mathbb{E}_y[||G(F(y)) - y||_1]. \tag{2.38}$$

In Equation 2.35, 2.36, 2.37 and 2.38, G is the mapping function from domain X to domain Y, D_Y is the discriminator to distinguish generated image $G(x)$ from real sample y. Besides, F is the mapping function from domain Y to domain X, D_X is the discriminator to distinguish generated image $F(y)$ from real sample x.

CycleGAN has the limited scalability that it can only learn the mapping between two domains. This requires multiple models to be trained in order to translate images across multiple domains. To address this problem, Choi *et al.* [42] proposed a unified generative adversarial networks (StarGAN) which allows us to learn the mapping between multiple domains with a single model. We describe details of the StarGAN [42] in Chapter 5.

Generative Adversarial Networks Approaches in ReID

Even though larger ReID datasets are available, the number of samples are still limited to train CNN models due to the expensive annotation process. Thus, over-fitting still can happen due to the lack of training samples in ReID dataset. To address this problem, some data augmentation methods have been proposed [37, 43, 134]. Zhong

et al. [134] proposed a random erasing technique which randomly selects the rectangle region and erases it with random values to avoid the over-fitting problem. In [37], DCGAN [131] was used to generate unlabeled person images. They also proposed the label smoothing regularization for outliers (LSRO) to assign the stable labels for the generated images. LSRP cam be formulated as

$$L_{LSRO} = -(1-\epsilon)\, log\, p(y) - \frac{\epsilon}{C}\sum_{c=1}^{C} log\, p(c), \tag{2.39}$$

$$q_{LSR}(c) = \begin{cases} 1-\epsilon+\frac{\epsilon}{C}, & c = y \\ \frac{\epsilon}{C}, & c \neq y \end{cases} \tag{2.40}$$

where $\epsilon \in [0, 1]$ is a hyper-parameter, C is the number of classes and $c \in 1, 2, ..., C$.

More recently, Zhong *et al.* [43] introduced a scene style transferred image generation using CycleGAN [133] as a data augmentation method for ReID. They also described improved label smoothing regularization (LSR) for generated images to address small portion of unreliable data. We will refer this method [43] as Camstyle for the rest of the thesis.

Zhang *et al.* propose a Crossing Generative Adversarial Network (Cross-GAN) [135] for learning a joint distribution for cross-image representations in an unsupervised manner. Cross-GAN employs the variational auto-encoder and adversarial layer to learn the latent representation. In [136], Similarity Preserving GAN (SimpGAN) is proposed to adopts the generative adversarial networks with the cycle consistency constraint to transform the unlabeled images from the source domain style to the target domain style. SimpGAN is used to model the cross-dataset domain style transfer in ReID.

In addition, Deng *et al.* [137] proposed a image-to-image translation model across different dataset domains while preserving self similarity and domain dissimilarity. Their method consists of an Siamese network and a CycleGAN. In [79], Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) was introduced to learn attribute-semantic and identity discriminative features simultaneously and transfer them to

the target domain without collecting additional data from the target domain. In order to reduce the dataset domain discrepancy, Wei *et al.* [138] proposed a Person Transfer Generative Adversarial Network (PTGAN) using a CycleGAN to learn the relationship between two different dataset domain.

CamStyle Adaptation for ReID [43]

CamStyle uses CycleGAN to generate scene style transferred images as extra training images. This method is to generate more images within one dataset domain but with different camera domains. Given a dataset with L different camera views, C_L^2 different CycleGAN models are separately trained for each pair combination of camera. To encourage the style-transfer to preserve the color consistency between generate image and the original image, identity mapping loss is added to the CycleGAN loss. The identity mapping loss is defined as:

$$V_{identity}(G, F) = \mathbb{E}_x[||F(x) - x||_1] + \mathbb{E}_y[||G(y) - y||_1] \tag{2.41}$$

where the variable notations are same as CycleGAN. All images are resized to 256x256. The generator has 9 residual blocks and four convolution layers. The discriminator follows 70x70 PatchGANs architecture [132].

2.2.5 Other Approaches

When the majority of clothing worn tends to be non-discriminative, ReID becomes very challenging. Attributes-based re-identification methods try to solve this problem by incorporating semantic attributes [139]. 'Jacket', 'female' and 'carried object' are all examples of semantic attributes. Layne *et al.* [70] has shown the feasibility of the semantic attributes for ReID. This method learns an attribute-specific, parts-based feature representation with attribute learning technique. Semantic attributes are mid-level features learned from a larger dataset a prior [140]. In [141], semantic attributes are combined with the low level features and is shown to improve the performance of ReID. In [142], CNN is used to learn a ReID embedding and predict

the person attributes simultaneously. This multi-task CNN learning integrates the person classification loss with the attribute classification loss.

3. IMPROVING VIDEO HEART RATE ESTIMATION

In this chapter, we describe the proposed method to improve existing video-based Heart Rate Estimation methods. We then describe the dataset that we used for the experiment and compare the performance with the Picard method and AFR method described in the Section 2.1.1.

3.1 Proposed Method

To mitigate described challenges in section 1.1, the present study employed small variation amplification, described in detail later, instead of ICA. Building on the strengths of previous approaches [65] we will only assess the green channel signal. Additionally, since HR is reflected in small color changes in skin region, using small variation amplification allows us to amplify the small color variations and attenuate large variations. To obtain a more stable estimation of the HR, we use a new approach to estimate the cutoff frequencies of the bandpass filters used before the Inter-Beat-Interval (IBI) computation (see Figure 2.1).

Figure 3.1 shows the overall system block diagram of the proposed method. Gray colored blocks indicate the extensions of the Picard [48] and AFR [53] methods mentioned in the section 2.1.1.

We first summarize the flow of the system and then describe the details of newly proposed blocks in the following subsections. Our proposed method begins with tracking the facial region on the subject to allow for HR estimates in both motion and no-motion conditions. Once the facial region is identified, then skin pixels are detected within the facial region. Face tracking and skin detection steps are distinct and are not used in the Picard and AFR. Details of face tracking and skin detection are given in the section 3.1.1. Next, the average of skin pixel intensities is computed

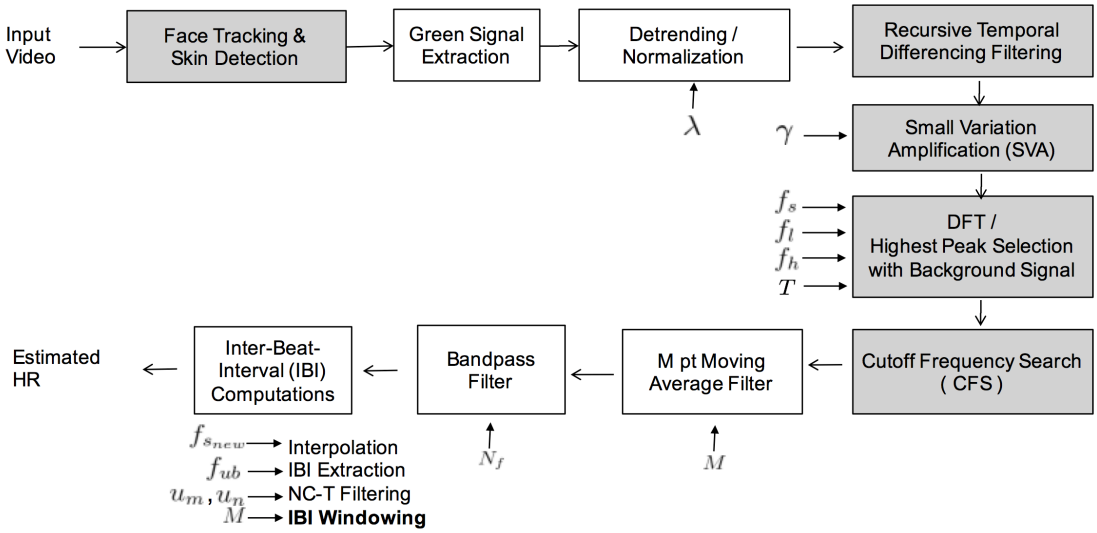


Fig. 3.1.: The overall block diagram of the proposed method.

in each frame to form the 1D signal. Unlike the use of RGB channel images in the Picard and AFR, we only use green channel image to form 1D signal. The reason why we use only the green channel image is that it has the most information relevant to HR. Kumar *et al.* [65] discussed that hemoglobin (Hb), which is related to blood volume change and eventually to HR, has a high absorption coefficient in the green wavelength. Detrending and normalization is the same as described in section 2.1.1.

Next temporal differencing filter and small variation amplification (SVA) is used instead of Independent Component Analysis (ICA). This is a new idea that amplifying the small color variations and attenuating the large variations in the skin regions to estimate HR. By using temporal differencing filter, we can get the difference information. We then use SVA to amplify the small difference and attenuate the large difference. Detailed formula is given in the following subsection 3.1.2.

As in AFR, we use the background (non-skin) signal but simply compare the highest peaks of the background and skin signals instead of frequency clustering. Power Spectrum Densities (PSDs) are obtained next for the amplified green signal from the skin and background regions. The highest peak is determined by comparing the highest peak frequencies from skin and background regions. The highest peak

is then used in the Cutoff Frequency Search (CFS) to find the cutoff frequencies for the bandpass filter. After a M-point moving average filter ($M = 5$), the signal is bandpass filtered with the cutoff frequencies from CFS.

Our method's IBI is almost same as the IBI of Picard and AFR. The only difference is the use of IBI windowing (boldface in Figure 3.1). IBI windowing uses a M point moving average filter ($M = 5$) on the NC-T filter output since ground truth HR from pulse oximeter is also moving window averaged.

3.1.1 Face Tracking and Skin Detection

The initial bounding box for tracking is obtained by face detection using a Haar Cascade Classifier [20]. For tracking, we derive a reference color model from the initial bounding box of the face region. For the color model, each RGB color space is quantized from the original 256 bins to 16 bins and is mapped into 1D 16^3-bin histogram. The sum of this histogram is then normalized to one. Particle filter tracking is used to find the corresponding face region in each frame [143]. Denoting the hidden state and the data at time t by x_t and y_t respectively, the probabilistic model we use for tracking is

$$p(x_{t+1}|y_{0:t+1}) \propto p(y_{t+1}|x_{t+1}) \int_{x_t} p(x_{t+1}|x_t)p(x_t|y_{0:t})dx_t \tag{3.1}$$

where $p(y_{t+1}|x_{t+1})$ is the likelihood model of the data, and $p(x_{t+1}|x_t)$ is the transition model of the second-order auto-regressive dynamics [143]. We define the state at time t as the location in the 2D image represented as pixel coordinates. For obtaining the likelihood $p(y_t|x_t)$, we use the distance metric $d(y) = \sqrt{1 - \rho\,(y)}$ where $\rho\,(y)$ is the sample estimate of the Bhattacharyya coefficient between the reference color model and the candidate color model of each particle at position y [144]. Face Tracking part is implemented and written by co-author Jeehyun Choe.

Given that our target signal includes only small color variations, even small levels of noise may significantly impact our HR estimates. Therefore, in an attempt to minimize noise, skin detection is use to remove hair, eye and mouth regions which

do not contain HR information. By using skin detection, we can also remove the eye-blinking artifact which is a large temporal variation.

A bayesian classifier using non-parametric density estimation is used for skin detection [145]. Let's assume that we have training images which are labeled for skin and non-skin pixels. The feature vector x_q is the quantized feature vector x by step Q [145].

$$x = (r, g, b) \rightarrow x_q = (r_q, g_q, b_q)$$

where (r,g,b) are pixel values in each color channel, $r_q = \lfloor \frac{r}{Q} \rfloor$, $g_q = \lfloor \frac{g}{Q} \rfloor$ and $b_q = \lfloor \frac{b}{Q} \rfloor$. We then construct the normalized histogram of quantized feature vector x_q for skin and non-skin pixels. The normalized histogram for each class is stored as a look-up table for the test classifications. Using the look-up table, we can classify the pixels values of new images which are not labeled. In Figure 3.2, white pixels indicate the skin pixels detected by a bayesian classifier within tracked face region. We only use the pixels which are denoted as white points for computing average to form 1D signal. As shown in Figure 3.2, we are able to remove the eyebrow or mouth regions which do not contain any HR signal by skin detection.

Fig. 3.2.: Example result of skin detection within tracked face.

3.1.2 Recursive Temporal Differencing Filter and Small Variation Amplification (SVA)

The basic idea is that we only amplify the small changes in time and suppress the large changes, because the HR signal is the small color changes in the skin region caused by cardiac activity. To achieve, a first order temporal recursive differencing filter is used on the detrended green channel signal:

$$g[n] = g[n+1] - g[n] \tag{3.2}$$

where $g[n]$ is the detrended green signal from skin pixels and n is the frame index.

Small variation amplification (SVA) is then used (Equations 3.3 and 3.4):

$$g_{amp}[n] = \alpha g[n] \tag{3.3}$$

$$\alpha = |g[n]|^{\gamma} \tag{3.4}$$

where α is the amplification factor based on the difference value of green signal. We choose $\gamma = -0.1$ empirically. From these blocks, we can suppress the large temporal variations in the signal and amplify the HR signal reflected in small temporal changes.

3.1.3 Peak Selection and Cutoff Frequency Search (CFS)

Peak Selection begins with finding the highest peak in the PSD for the skin and background regions. If the highest peak from the skin region is similar to the highest peak from the background region, then this is a strong periodic noise signal from factors such as lighting. The similarity between highest peaks is determined by:

$$d_f = |f_1 - f_2| \tag{3.5}$$

where f_1 is the highest peak frequency from the skin region and f_2 is the highest peak frequency from the background region. If $d_f < T$, we then find the next highest peak in the skin PSD. We empirically choose threshold $T = 0.1$. The selected highest peak

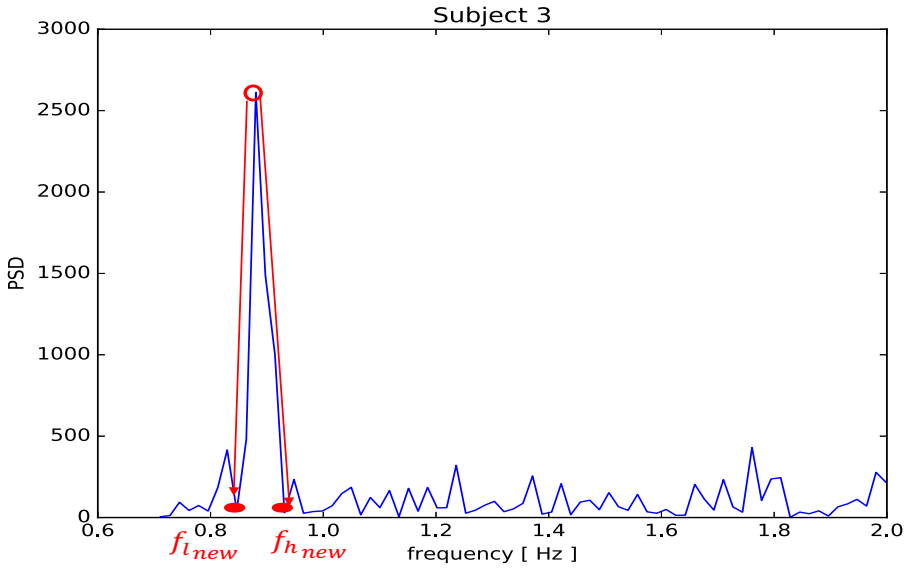


Fig. 3.3.: An example of CFS in PSD.

is used as the starting point for the CFS which eventually determines tighter cutoff frequencies for the Bandpass Filter (BPF).

As shown in Figure 3.3, a gradient search is done on the skin region PSD domain starting from the highest peak. The points that have a sign change are determined as the new cutoff frequencies ($f_{l_{new}}$ and $f_{h_{new}}$) for BPF. Tighter cutoff frequencies were achieved using CFS for BPF and this supports more stable estimations in the IBI computation.

3.2 Experimental Results

3.2.1 Experimental Setup and Dataset

In our experiments, we acquired two different datasets. Both were collected in the same room which had windows with semi-transparent blinds and lighting on the ceiling as shown in Figure 3.4. Camera 1 in Figure 3.4 was used to record the subject and camera 2 was used to record the pulse oximeter for the ground truth HR. All videos had a spatial resolution of 1920×1080, 30 fps and 60 seconds length.

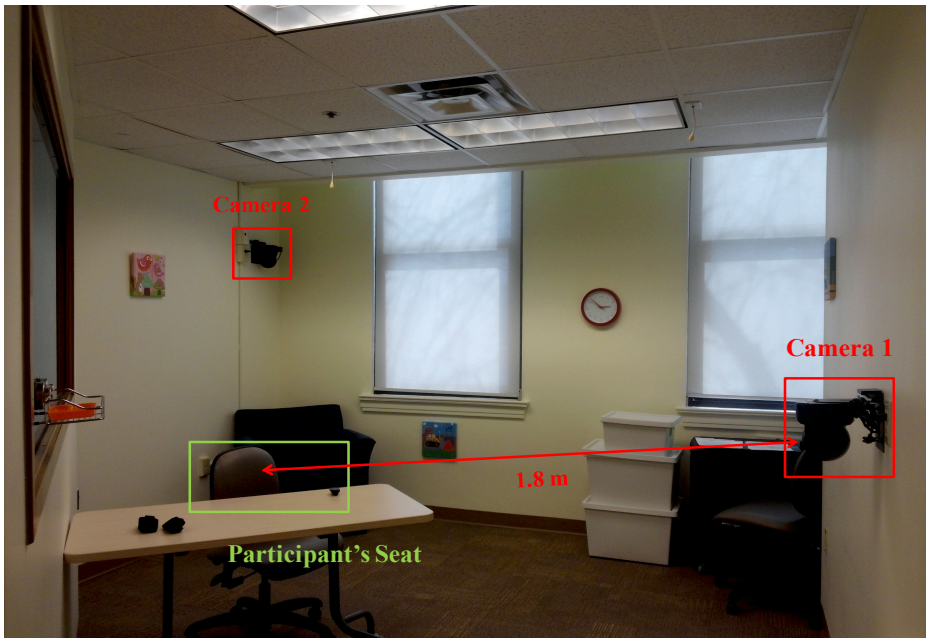


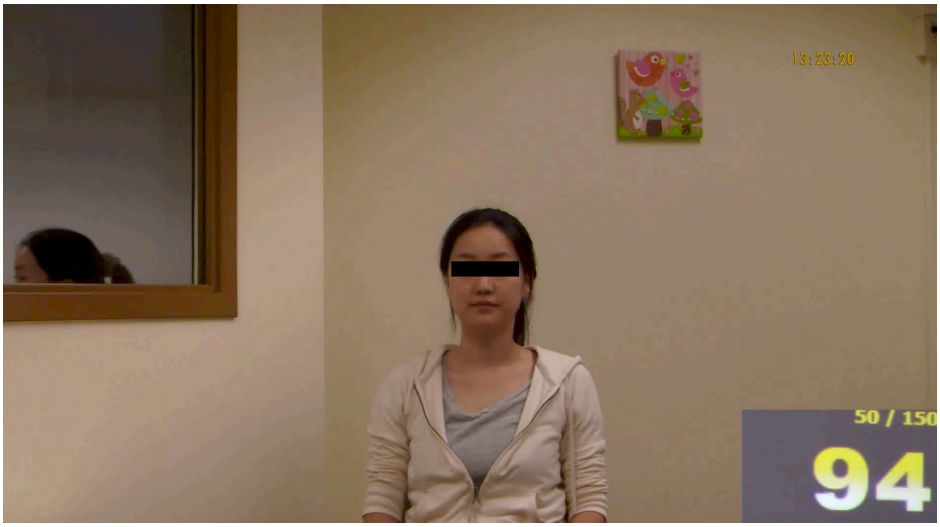
Fig. 3.4.: The room and camera setting.

The ground truth HR was measured using a Pulse Oximeter attached to the finger of the subject. A Nonin GO_2 Achieve Finger Pulse Oximeter was used in Dataset 1 and CE & FDA Approved Handheld Pulse Oximeter was used in Dataset 2. The pulse oximeter HRs were manually recorded from the video once per second in both datasets. Dataset 1 included 22 subjects (12 females and 10 males) and Dataset 2 included 18 subjects (9 females and 9 males).

The distance between the subject and the camera was approximately 1.8 m in both datasets, the zoom was manually adjusted to focus only on the upper torso and face in Dataset 1. In Dataset 2, the zoom was manually adjusted to show entire upper body of the subject. Examples of videos from Dataset 1 and 2 are provided in Figure 3.5.

Dataset 1 included no-motion videos and Dataset 2 included both non-random motion and no-motion videos. In the no-motion videos, the subjects were seated and were asked to look toward the camera. In the non-random motion videos, the subjects were asked to move their head from left to right repeatedly while keeping their faces

(a) Dataset 1 Video Setting.



(b) Dataset 2 Video Setting.

Fig. 3.5.: Video Setting Examples.

toward the camera. I In our experiments, we set the parameters for Picard [48] and AFR [53] as described in the section 2.1.1. A summary of the parameters used in the three methods are provided in Table 3.1. All the parameters are chosen empirically except the video frame rate f_s.

The initial facial region was detected in the first frame of the video [20]. 80% of the detected face region's width and height were used with all three methods. Skin

Table 3.1.: A summary of parameters used in the three VHR methods.

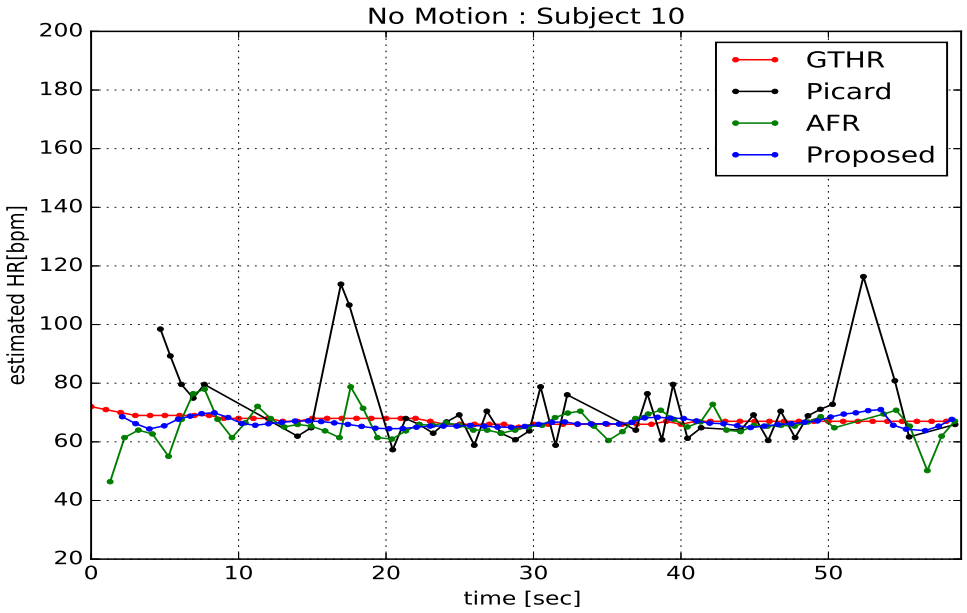
Parameter	Picard [48]	AFR [53]	Proposed
λ	100		
f_s	30		
(f_l,f_h)	(0.7,2)		
$f_{s_{new}}$	256		
M	5		
N_f	127		
(u_n, u_m)	(0.4,1)		
t_r	-	0.1	-
t_n	-	0.1	-
t_m	-	0.4	-
γ	-	-	-0.1
T	-	-	0.1

detection was trained from the SFA image database [146]. The background region is selected on the wall behind the subject (with the size of 80% of face region).

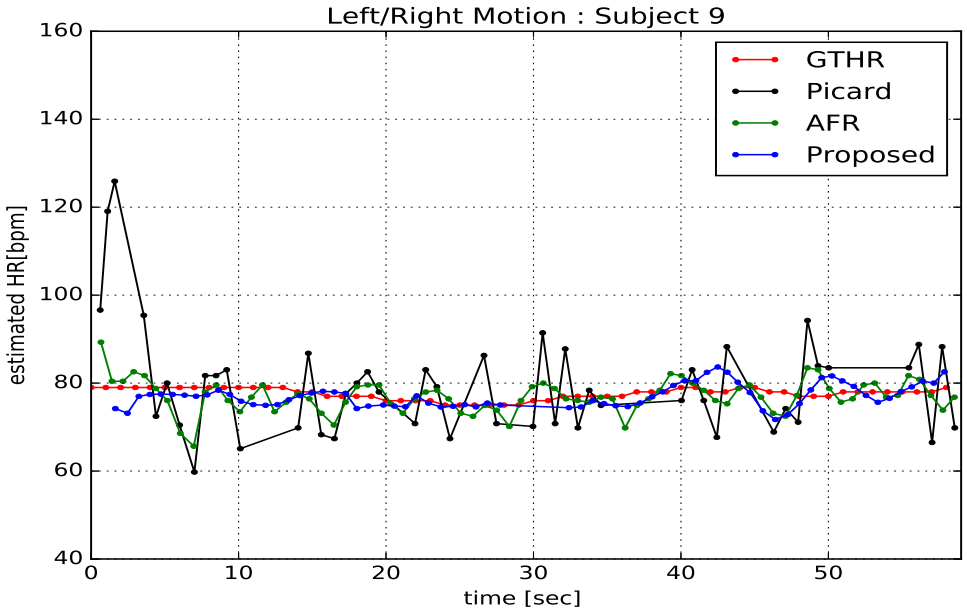
3.2.2 Error Rate Comparison

The proposed method shows better overall performance, compared to the Picard [48] and AFR [53]. For example, Figure 3.6 shows one of the final HR estimation from all three methods for the motion video of Subject 9 in Dataset 2. The red line shows manually annotated ground truth HR from the pulse oximeter. The estimated HR from Picard, AFR and the proposed method are labeled as black, green, and blue respectively. Our proposed method shows the closest estimation to true HR in this case.

To evaluate the overall performance, we defined the error metrics as:



(a) Subject 10 with no-motion in Dataset1.



(b) Subject 9 with motion in Dataset2.

Fig. 3.6.: A comparison of the three VHR methods : Estimated HR [bpm] vs Time [sec].

$$e_1 = \begin{cases} \frac{1}{N}\sum_{n=1}^{N} |\frac{HR_{est}[n] - HR_{true}[n]}{HR_{true}[n]}| * 100 \quad [\%] \\ \\ \frac{1}{N}\sum_{n=1}^{N} |HR_{est}[n] - HR_{true}[n]| \quad [bpm] \end{cases} \tag{3.6}$$

where $HR_{est}[n]$ is the estimated HR in units of beat per minute (bpm), $HR_{true}[n]$ is the manually recorded HR in units of bpm from the pulse oximeter and n is the time domain index when the HR estimations exist. Comparison results for the three VHR methods in terms of Error Rate e_1 in both units [bpm and %] are shown in Table 3.2, 3.3 and 3.4.

In Dataset 1, the proposed method has the lowest average error rate across all 22 subjects (3.67 % and 4.71 bpm) as shown in Table 3.2. The proposed method outperforms Picard and AFR in most of the videos except three cases. The overall error rate in Dataset 1 is lower than the one in Dataset 2 because Dataset 1 has more pixels in the face region due to the manual zoom focus on the upper torso and face (compared to the whole upper body in Dataset 2).

In Dataset 2, the proposed method has the lowest average error rate across all 18 no-motion videos (7.09 % and 9.48 bpm) as shown in Table 3.3. The proposed method outperforms Picard and AFR in most of the videos except four cases. Even though the proposed method outperforms the previous methods in Dataset 2, the overall error is still higher than Dataset 1 due to the less number of pixels in face region.

In non-random motion videos, all three methods still have a large error rate compared to the no-motion scenario. The proposed method has the lowest average error rate as measured by percent across all 18 subjects (16.89 %), while the Picard method has the lowest average error rate as measured by bpm across all 18 subjects (13.36 bpm) as shown in Table 3.4. The proposed method outperforms Picard and AFR in most of the videos except five cases. However, when the highest peak is not matched

Table 3.2.: A comparison of three VHR methods in Dataset 1 with no motion : Error Rate e_1 [bpm , %]

ID	Picard [48]		AFR [53]		Proposed	
	[bpm]	[%]	[bpm]	[%]	[bpm]	[%]
1	5.53	8.00	1.49	2.16	1.58	2.29
2	3.58	6.32	2.97	5.27	1.81	3.21
3	5.50	10.44	1.89	3.59	1.19	2.27
4	6.47	7.23	2.75	3.08	2.44	2.73
5	11.29	21.21	2.69	5.03	2.26	4.21
6	12.37	13.67	4.47	4.92	3.15	3.46
7	8.36	10.66	2.89	3.69	2.36	3.01
8	7.12	10.00	2.95	4.12	1.93	2.69
9	5.95	7.91	2.39	3.17	2.10	2.78
10	6.57	6.99	2.85	3.03	15.15	16.05
11	5.03	6.37	1.43	1.82	1.91	2.44
12	5.34	9.82	4.28	7.82	2.47	4.49
13	16.49	22.11	16.25	21.16	14.02	18.17
14	5.77	7.26	2.88	3.63	1.84	2.32
15	9.29	12.96	5.05	6.88	4.15	5.65
16	8.56	12.25	6.58	9.35	3.03	4.33
17	4.38	6.55	2.76	4.12	2.29	3.44
18	4.79	6.32	1.73	2.29	1.23	1.62
19	8.00	8.80	2.94	3.21	1.54	1.68
20	6.24	7.52	2.58	3.11	1.96	2.37
21	9.87	11.24	6.19	7.04	8.84	9.74
22	7.98	10.33	4.00	5.17	3.57	4.63
Avg.	7.48	10.18	3.82	5.17	**3.67**	**4.71**

Table 3.3.: A comparison of three VHR methods in Dataset 2 with no motion: Error Rate e_1 [bpm , %]

ID	Picard [48]		AFR [53]		Proposed	
	[bpm]	[%]	[bpm]	[%]	[bpm]	[%]
1	21.44	32.28	13.06	20.04	6.91	10.51
2	10.90	15.38	7.62	10.58	6.50	9.12
3	10.67	13.10	9.89	12.15	9.58	11.75
4	19.40	26.85	3.86	5.29	36.96	50.81
5	10.11	10.44	5.93	6.09	6.13	6.33
6	15.94	21.34	14.84	19.88	2.49	3.32
7	8.31	10.04	3.40	4.10	3.71	4.46
8	16.39	23.00	10.31	14.30	7.70	10.65
9	11.17	14.35	9.09	11.69	5.85	7.53
10	9.29	13.78	3.83	5.64	1.46	2.17
11	17.41	19.07	28.60	31.39	8.13	8.91
12	12.70	18.55	8.15	11.79	4.24	6.11
13	15.24	16.86	23.02	25.54	3.26	3.60
14	7.46	8.84	5.06	6.01	4.07	4.85
15	7.99	11.08	4.66	6.48	3.02	4.16
16	8.99	11.83	5.94	7.87	6.66	8.86
17	20.05	38.98	6.22	11.16	6.09	10.87
18	6.49	9.14	4.87	6.75	4.77	6.59
Avg.	12.77	17.50	9.35	12.04	**7.09**	**9.48**

Table 3.4.: A comparison of three VHR methods in Dataset 2 with non-random motion: Error Rate e_1 [bpm , %]

ID	Picard [48]		AFR [53]		Proposed	
	[bpm]	[%]	[bpm]	[%]	[bpm]	[%]
1	10.86	14.38	9.92	13.13	4.49	5.94
2	20.78	29.57	20.54	29.27	6.82	9.68
3	13.22	16.99	11.27	14.87	5.88	7.81
4	14.56	18.86	15.53	19.89	27.74	35.65
5	19.52	20.55	27.39	28.93	34.08	36.01
6	23.98	26.35	31.15	34.23	18.81	20.67
7	16.34	16.71	17.83	18.18	14.62	14.88
8	10.73	15.12	15.00	21.03	11.28	15.79
9	8.96	11.49	2.66	3.42	2.08	2.66
10	9.70	13.32	9.68	13.30	20.85	28.56
11	14.82	15.96	14.82	15.96	21.33	22.95
12	7.79	11.79	4.74	7.19	2.76	4.17
13	13.29	15.38	35.47	40.69	25.20	28.90
14	12.59	15.68	14.63	18.24	26.58	33.08
15	13.87	23.01	11.75	19.36	11.24	18.53
16	8.33	10.35	4.92	6.08	3.57	4.44
17	11.75	19.75	6.44	10.78	2.64	4.42
18	9.39	11.76	8.20	9.83	8.14	9.89
Avg.	**13.36**	17.06	14.55	18.02	13.78	**16.89**

with the true HR, the proposed method has the tendency to propagate the error because of CFS.

3.2.3 Statistical Analyses

Table 3.5.: No-motion videos, Number of Sample : 40.

Pair	Methods	Mean	SD	p-value
1	Proposed	6.8545	8.13569	
	AFR	8.2603	6.79649	0.365
2	Proposed	6.8545	8.13569	
	Picard	13.4718	7.35702	< 0.05
3	AFR	8.2603	6.79649	
	Picard	13.4718	7.35702	< 0.05

Table 3.6.: Non-random motion videos, Number of Sample : 18.

Pair	Methods	Mean	SD	p-value
1	Proposed	16.8906	11.62436	
	AFR	18.0211	9.96929	0.635
2	Proposed	16.8906	11.62436	
	Picard	17.0567	5.22310	0.952
3	AFR	18.0211	9.96929	
	Picard	17.0567	5.22310	0.602

We used a series of the paired samples t-tests [147] to compare the proposed method to the other two methods. The outcome variable was the percentage error rate $e_1[\%]$ for no motion and non-random motion videos. Two-tailed tests were used

with an alpha level of 0.05. Descriptive statistics, including mean error rate (M) and standard deviation (SD) are presented in Table 3.5 and 3.6.

For the no-motion videos (N = 40, Table 3.5), the proposed method had a lower mean error rate (M) than both other methods, but was only significantly lower than the Picard method ($p < 0.05$), not AFR ($p = 0.37$). The AFR error was also significantly lower than Picard ($p < 0.05$). Thus both the proposed and AFR methods outperformed the Picard method. For the non-random motion videos (N = 18, 3.6), the same pattern of results emerged; however, method differences were not statistically significantly different.

3.3 Spatial Pruning

We have previously proposed the usage of skin detection from the section 3.1. However, even though the skin detection performs pretty well, it still heavily depends on the illumination changes, scene changes or subject's skin tone color in the videos. If skin detection produces a lot of error in our system, it will produce critical noise in our target signal. Thus, we propose a new method called spatial pruning as an alternative method of skin detection. The idea here is that we have already tracked face region and that face region is composed of skin region mostly. So if we can cluster the pixels into two different groups based on the distribution of pixel intensities, we then can separate only skin pixels from the tracked face.

We have two assumptions in this method:

1. Histogram of face region has bimodal distribution.

2. The higher bump in bimodal distribution represents pixel value and the lower bump represents the other noise regions such as mouth, eyebrow or small background in the edge.

Figure 3.7 shows the overall flow of the proposed spatial pruning. Input of this block is the tracked face region. Within the face region, we can compute the histogram

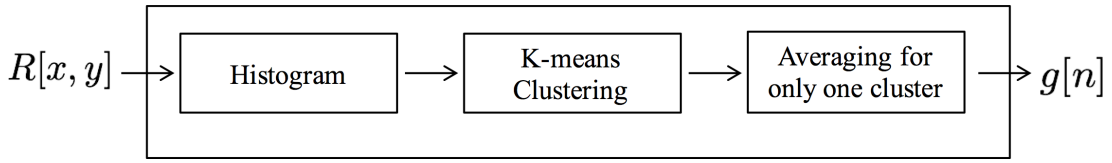


Fig. 3.7.: The block diagram of the spatial pruning.

which represents the distribution of pixels. Histogram of the face region is computed by using Equation 3.7

$$h_k = \frac{N_k}{N} \tag{3.7}$$

where $k = 0, 1, \ldots, 255$ is the bin index, N_k is the number of pixels which have intensity value k and N is the total number of pixel. This histogram h_k is fed into K-means clustering to separate two clusters (one for skin, one for nose). Based on our assumptions described above, we choose one cluster which contains the most frequent pixel value. The chosen cluster's values are averaged to obtain the output signal $g[n]$.

3.3.1 K-means Clustering

In this section, we introduce the theory and details of K-means Clustering method. K-means Clustering is a clustering technique that attempts to find K number of clusters, which are represented by their centroids [148]. We first choose K initial centroids (K is a user determined parameter). In our case, K = 2 since we want to cluster into skin or noise cluster.

1. Select randomly K points for initial centroids.

2. Form K clusters by assigning each data point to its closest centroid. To determine the closest centroid, Euclidean distance is typically used. This step can

be explained as minimizing the objective function Sum of the Squared Error (SSE) 3.8

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} ||c_i - x||^2 \tag{3.8}$$

where C_i is the ith cluster, c_i is the cluster centroid, K is the total number of clusters, $|| * ||^2$ is the squared euclidean distance, and x is a data point (a histogram value in this application).

3. Update the new centroid for each cluster by using Equation 3.9

$$c_i = \frac{1}{N_i} \sum_{x \in C_i} x \tag{3.9}$$

where N_i is the number of data points in the ith cluster and the rest of the notations follows the Equation 3.8.

4. Repeat step 1, 2 until centroids do not change. Since K-means Clustering is an iterative method, we need to give the number of iteration to stop the search. We chose empirically 100 iterations in this experiment.

3.3.2 Preliminary Result and Discussion

To check the feasibility of the idea, we have done the preliminary test on the real data. Figure 3.8 shows the clustered histogram using K-means clustering on the tracked face region. As in the first assumption of section 3.3, face histogram shows the bimodal distribution. We use the K-means clustering (in our case, K = 2) to separate the skin region and noise region. Since the histogram has bimodal distribution (clearly separable data points), K-means clustering method successfully separates two clusters in the histogram. Red color denotes the first cluster and blue color denotes the second cluster in Figure 3.8.

Based on our second assumption in section 3.3, red cluster will be skin cluster and blue cluster will be noise cluster. To verify our second assumption, we have done rough inver-mapping to the frame from the skin cluster values by using Equation 3.10

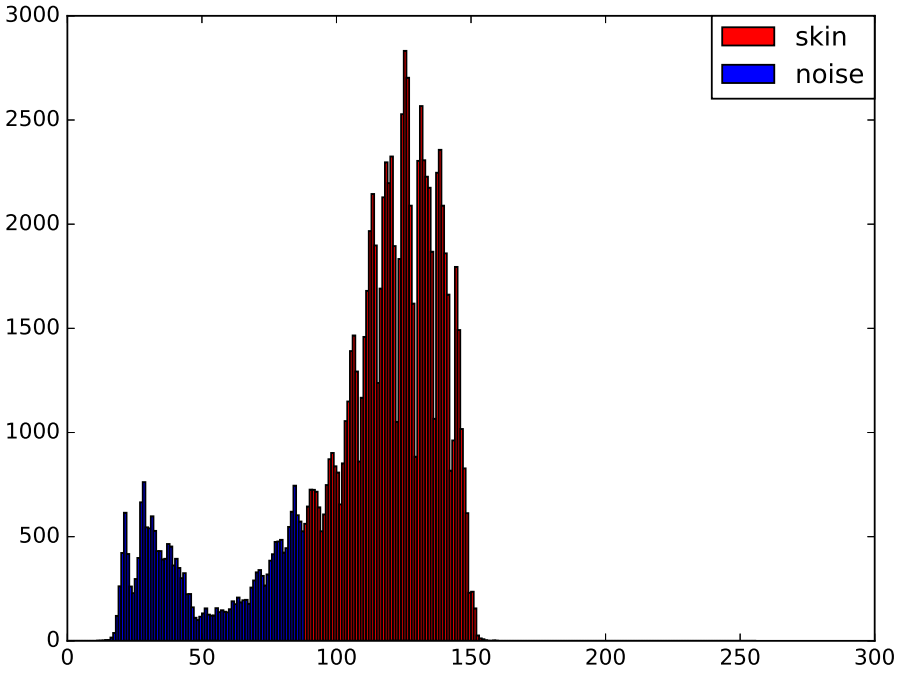
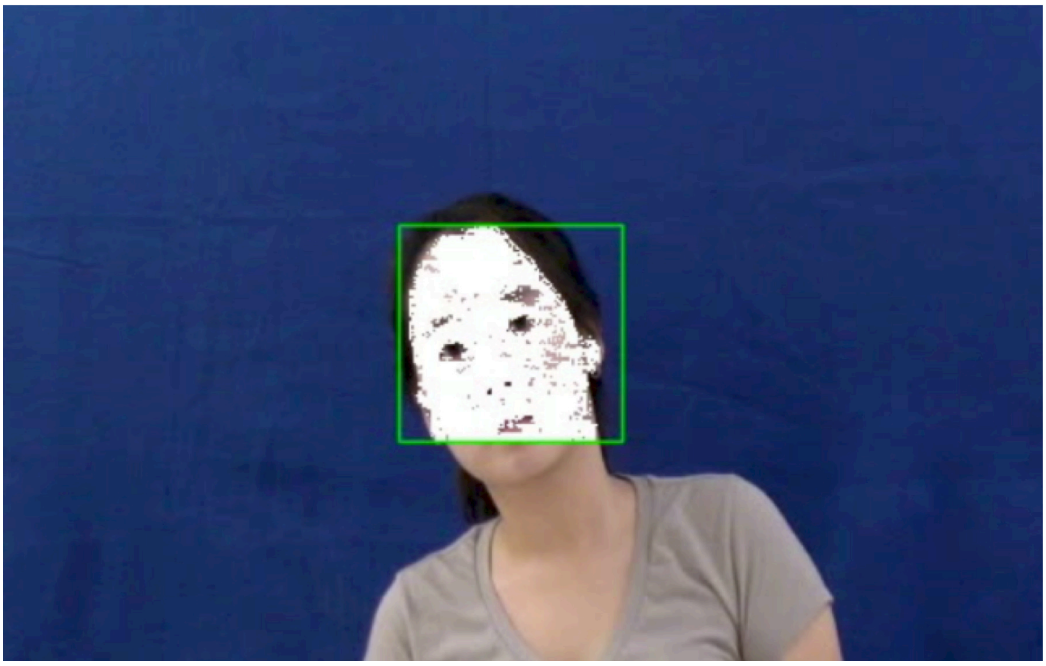


Fig. 3.8.: Clustered Histogram using K-means clustering (K=2).

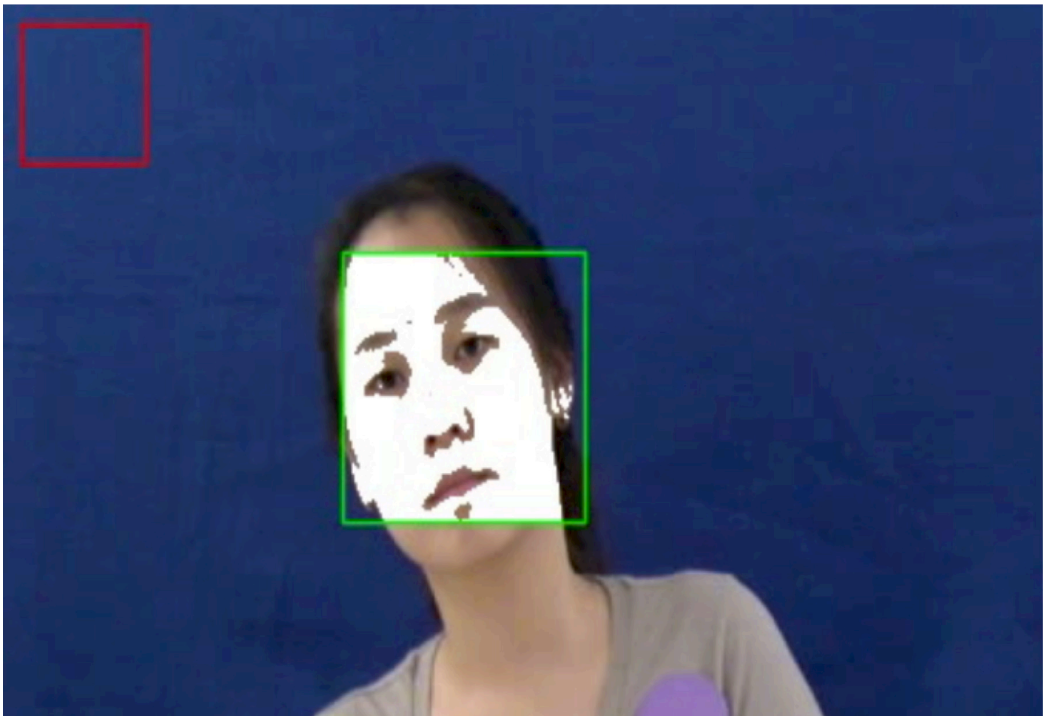
$$L^{'}[x,y] = \begin{cases} 0, & \text{if } |L[x,y] - h_{max}| < 30 \\ L[x,y], & \text{otherwise} \end{cases} \tag{3.10}$$

where $L[x,y]$ is the pixel intensity at $[x,y]$ and $h_{max} = \underset{k}{\operatorname{argmax}}\, h_k$ (the histogram bin which has the maximum occurrence). The result shown in Figure 3.9b verifies that our second assumption is right. White pixels denote pixels classified as skin in Figure 3.9a and 3.9b.

Even though we manually choose the skin detection parameter (which is not feasible in practice) for the maximum performance in this experiment, it has many false negative error around cheek area as shown in Figure 3.9a. Compare to skin detection, our approach shows more stable and well-classified skin region even though we only



(a) Skin Detection Sample Result.



(b) Clustering Preliminary Result.

Fig. 3.9.: A Comparison of Skin Detection and Spatial Pruning.

use rough approximation method for preliminary experiment as shown in Equation 3.10.

Since the preliminary result is promising for this approach, we plan to elaborate the idea with mathematical background and run the experiment with entire dataset. For example, since we have two clusters from the method, we model two clusters defined by the following Equations 3.11 and 3.12

$$\mu_i = \frac{1}{N_i} \sum_{L[x,y] \in C_i} L[x, y] \tag{3.11}$$

$$\sigma_i{}^2 = \frac{1}{N_i - 1} \sum_{L[x,y] \in C_i} (L[x, y] - \mu_s)^2 \tag{3.12}$$

where i is the cluster index (i ={skin,noise}), C_i is the i-th cluster, N_i is the number of points in the C_i and μ_i, σ_i are mean and variance of C_i respectively.

Let C_1 be the skin cluster. Instead of using entire points in C_1, we are planning to model skin pixel distribution as gaussian distribution with μ_1, σ_1 and use the points only distributed near μ_1.

3.4 Motion Artifact Modeling based on Illumination

Since our proposed method in section 3.1 did not provide any statistically significant improvement in non-random motion videos, we model in this section 1D green signal to analyze existing noise in HR signal. Our mathematical model is based on motion noise caused by illumination changes on the skin surface.

When an image is obtained by a camera, the measured intensity of an image is the product of incident illumination and reflectance at a surface [149]. (Equation 3.13)

$$L[x, y, n] = I[x, y, n] \cdot R[x, y, n] \tag{3.13}$$

In Equation 3.13, $L[x, y, n]$ is the intensity of the image, $I[x, y, n]$ is the intensity of incident illumination and $R[x, y, n]$ is the reflectivity of the surface. n is the frame index and $[x, y]$ is the two-dimensional spatial coordinate in n-th frame.

Intensity of incident illumination $I[x, y, n]$ can be expressed by three factors [149] as shown in Equation 3.14

$$I[x, y, n] = I_0 + I_s \cdot cos\big[\theta[x, y, n]\big] \tag{3.14}$$

where I_0 is the intensity of uniform diffuse illumination (e.g., background light, sunlight) and I_s is the intensity of the point light source. We can assume that I_0, I_s are constants since the illumination source is not variant in our case. $\theta[x, y, n]$ is the angle between incidence illumination and surface normal. As the subject is moving in video, the incident angle $\theta[x, y, n]$ will vary along the motion. We should note that most of the VHR methods [48] [53] [65] assume the constant illumination. In this model, we assume that illumination changes in function of the incident angle ($\theta[x, y, n]$).

The reflectance $R[x, y, n]$ at a surface is the intensity of the light reflected back from skin. The light reflected back from the skin consists of two different levels of reflectance based on the skin optical model [150]: (i) surface reflectance and (ii) subsurface reflectance or backscattering. A large portion of light is reflected by the skin surface. Remaining part of the light is reflected by subsurface. This subsurface reflectance contains the heart rate information because it reflects the blood volume change by chromophores (HbO_2) [65]. From skin optical model shown in Figure 1.1, the reflectance signal can be modeled as Equation 3.15

$$\begin{aligned} R[x, y, n] &= a[x, y, n] \cdot h[n] + b[x, y, n] \\ &\simeq a \cdot h[n] + b \end{aligned} \tag{3.15}$$

where a is the strength of blood perfusion and b is the surface reflectance. a, b are assumed to be constants in the same subject because it is a characteristic of the human subject. We can also assume that $b >> a$ since b is the intensity of a large portion of reflected light from light source.

Using Equation 3.14 and 3.15, we model the intensity of the image by Equation 3.16.

$$\begin{aligned} L[x,y,n] &= \big(I_0 + I_s \cdot cos\big[\theta[x,y,n]\big]\big)\big(a \cdot h[n] + b\big) \\ &= I_0 \cdot a \cdot h[n] + I_s \cdot a \cdot cos\big[\theta[x,y,n]\big] \cdot h[n] \\ &\quad + I_s \cdot b \cdot cos\big[\theta[x,y,n]\big] + I_0 \cdot b \end{aligned} \tag{3.16}$$

Equation 3.17 models 1D green signal of the method in chapter 3.1 by averaging over skin region pixels within the frame

$$\begin{aligned} g[n] &= \frac{1}{N_s} \sum_{x \in S_x} \sum_{y \in S_y} L[x,y,n] \\ &= \frac{1}{N_s} \sum_{x \in S_x} \sum_{y \in S_y} \Big(I_0 \cdot a \cdot h[n] + I_s \cdot a \cdot cos\big[\theta[x,y,n]\big] \cdot h[n] \\ &\qquad + I_s \cdot b \cdot cos\big[\theta[x,y,n]\big] + I_0 \cdot b \Big) \end{aligned} \tag{3.17}$$

where S_x , S_y are the skin pixel coordinates and N_s is the total number of skin pixel in one frame. To ease the notation, let time varying term irrelevant to HR signal be $m[n]$.

$$m[n] = \frac{1}{N_s} \sum_{x \in S_x} \sum_{y \in S_y} cos\big[\theta[x,y,n]]\big] \tag{3.18}$$

Then, $g[n]$ can be simplified using Equation 3.18, $A = I_0 \cdot a$, $B = I_s \cdot a$, $C = I_s \cdot b$ and $D = I_0 \cdot b$ in Equation 3.19.

$$\begin{aligned} g[n] &= I_0 \cdot a \cdot h[n] + I_s \cdot a \cdot m[n] \cdot h[n] + I_s \cdot b \cdot m[n] + I_0 \cdot b \\ &= A \cdot h[n] + (B \cdot h[n] + C) \cdot m[n] + D \end{aligned} \tag{3.19}$$

Note that $m[n]$ is the motion artifact signal since $\theta[x,y,n]$ varies along the subject's motion. In the final form of 1D raw green signal $g[n]$, there are multiplicative motion artifact term and also additive motion artifact term.

In order to analyze the motion artifact effect in frequency domain, we compute PSD of $g[n]$ and it can be expressed in terms of PSD of $h[n]$ and $m[n]$ as shown in Equation 3.20

$$\big|G[k]\big|^2 = A^2\big|H[k]\big|^2 + B^2\big|H[k] * M[k]\big|^2 + C^2\big|M[k]\big|^2 + D^2\delta[k] \tag{3.20}$$

where $H[k]$, $M[k]$ here are Discrete Fourier Transform(DFT) coefficients of $h[n]$, $m[n]$ respectively. As shown in Equation 3.20, $G[k]$ has still the heart rate signal $H[k]$ but it is involved with the convolution term. Based on the motion artifact frequencies in $M[k]$, the convolution term will produce noisy signal near $H[k]$ since $H[k]$ will be shifted by factor of $M[k]$'s frequencies. This noisy signal in PSD can overwhelm the HR signal based on the constant values (A,B,C,D). If the subject has low level of blood perfusion (small a), HR signal will be significantly overwhelmed by motion artifact signal.

To verify this model, we test on the synthetically generated data. Let $h[n]$ as the heart rate signal which is periodic signal given by Equation 3.21

$$h[n] = cos(2\pi f_{hr} n) \tag{3.21}$$

where $f_{hr} = 1$ Hz (60 bpm) is heart beat frequency. Motion artifact factor $m[n]$ is given as summation of two different frequencies f_1 and f_2

$$m[n] = cos(2\pi f_1 n) + cos(2\pi f_2 n) \tag{3.22}$$

where $f_1 = 0.25$ Hz, $f_2 = 0.15$ Hz. The constants are empirically chosen based on the skin optics principles : $I_0 = 1$, $I_s = 3$, $a = 5$, $b = 50$. By substituting the set values to the Equation 3.20, we then can derive the final formula of $G[k]$ as given in Equation 3.25

$$\begin{aligned} H[k] * M[k] = \frac{1}{4}[&\delta[k-1.25] + \delta[k-1.15] + \delta[k-0.85] + \delta[k-0.75] \\ &+ \delta[k+1.25] + \delta[k+1.15] + \delta[k+0.85] + \delta[k+0.75]] \end{aligned} \tag{3.23}$$

$$\begin{aligned} G[k] &= 5H[k] + 15H[k] * M[k] + 150M[k] + 50\delta[k] \\ &= 50\delta[k] \ + 2.5\delta[k-1] + 2.5\delta[k+1] \\ &\quad + 75\big[\delta[k-0.25] + \delta[k+0.25] + \delta[k-0.15] + \delta[k+0.15]\big] \\ &\quad + 3.75\big[\delta[k-1.25] + \delta[k-1.15] + \delta[k-0.85] + \delta[k-0.75] \\ &\quad + \delta[k+1.25] + \delta[k+1.15] + \delta[k+0.85] + \delta[k+0.75]]\big] \end{aligned} \tag{3.24}$$

$$
\begin{aligned}
G(f) &= 5H(f) + 15H(f) * M(f) + 150M(f) + 50\delta(f) \\
&= 2.5\delta(f-1) + 2.5\delta(f+1) \\
&\quad + 3.75\big(\delta(f-1.25) + \delta(f-1.15) + \delta(f-0.85) + \delta(f-0.75) \\
&\quad + \delta(f+1.25) + \delta(f+1.15) + \delta(f+0.85) + \delta(f+0.75)\big) \\
&\quad + 75\big(\delta(f-0.25) + \delta(f+0.25) + \delta(f-0.15) + \delta(f+0.15)\big) \\
&\quad + 50\delta(f)
\end{aligned}
\tag{3.25}
$$

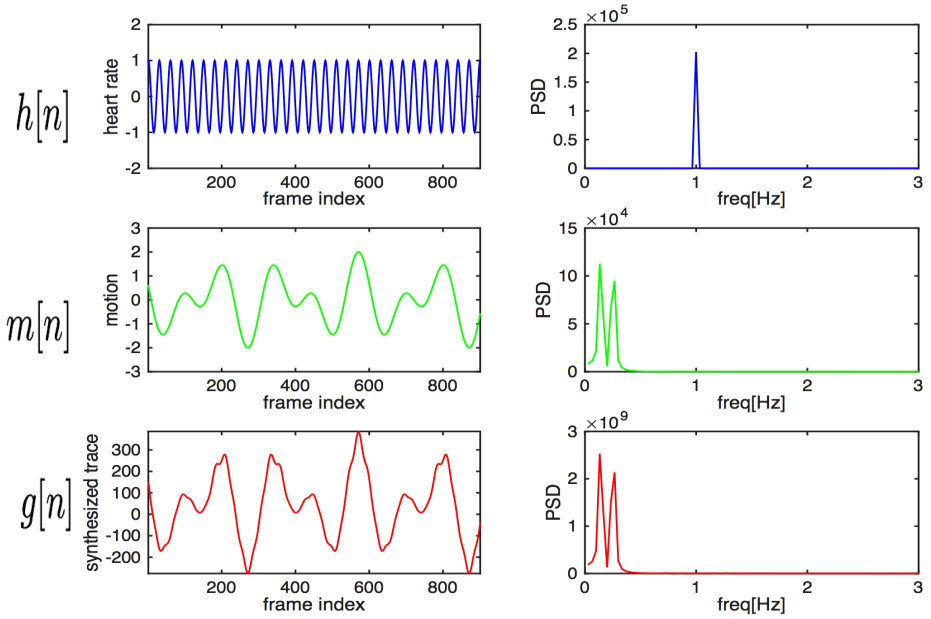
Figure 3.10 shows the time domain signal and power spectrum density of each synthetic data $h[n]$, $m[n]$ and $g[n]$. Figure 3.10a demonstrates the broad range of PSD to see the low frequency motion effects. Figure 3.10b shows the PSD within 0.7 - 3 Hz only. As shown in that, HR signal (1 Hz signal) has been overwhelmed by motion artifact terms because of modulation part in Equation 3.25. Thus, we have shown that HR signal can be overwhelmed by the motion artifact terms caused by illumination incident angle changes.

3.5 Separation of Illumination and Reflectance using Homomorphic Filter

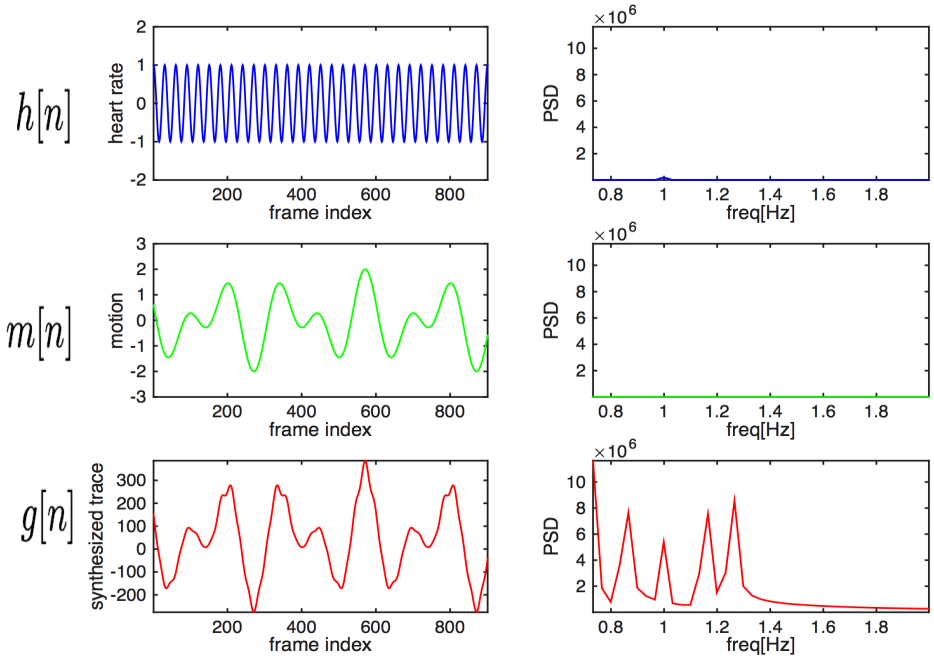
Based on our motion artifact modeling in the previous section, we define the motion artifact problem as illumination incident changes. Thus we propose a separation of Illumination and Reflectance from the given image by using homomorphic filter. According to the Equation 3.13, Illumination $L[x, y, n]$ and Reflectance $R[x, y, n]$ has multiplicative relation. General approach of separating multiplicative model in signal processing starts with taking logarithm [151]. By taking logarithm, the result will be additive modeling as shown in Equation 3.26

$$
log(L[x, y, n]) = log(I[x, y, n]) + log(R[x, y, n]) \tag{3.26}
$$

Based on the typical usage of Homomorphic Filter for separating illumination and reflectance [152], we propose a new block shown in Figure 3.11.



(a) Full Power Spectrum Density.



(b) Partial Power Spectrum Density (0.7 - 3 Hz).

Fig. 3.10.: Synthetic Data Simulation Result.

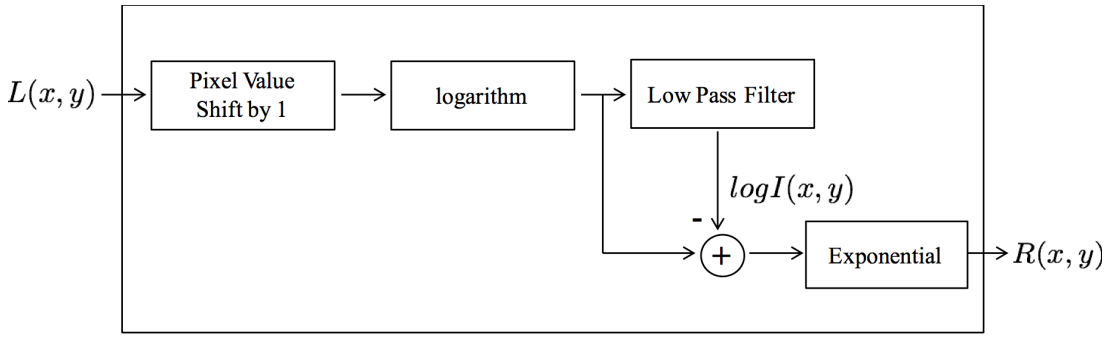


Fig. 3.11.: The block diagram of the homomorphic filter.

It begins with taking logarithm. Based on the fact that illumination is slowly changing component (low frequency component), we take low pass filtering on the image to estimate illumination $log(\hat{I}[x,y,n])$. By subtracting estimated illumination from $log(L[x,y,n])$, we can estimate Reflectance $R[x,y,n]$ as shown in Equation 3.27

$$log(\hat{R}[x,y,n]) = log(L[x,y,n]) - log(\hat{I}[x,y,n]) \tag{3.27}$$

Finally, we take exponential back to go back to original coordinate from the log coordinate. Thus, if we can separate successfully the reflectance from the given image, we will have illumination free information for future HR estimation. In the end, we expect more accurate HR estimation by employing this idea.

4. A TWO STREAM SIAMESE CNN FOR PERSON RE-IDENTIFICATION

Zheng *et al.* [76] demonstrates that multi-shot scenario is superior to single-shot scenario using empirical evidences. Their experiments show that multi-shot approaches are more favorable since both probe and gallery contain much richer visual information as compared to single image. In addition, combining spatial features using multi-shot helps address the challenges associated with viewpoint and pose invariance [36, 153]. Also, in real-life surveillance systems, human detection and tracking methods generate multiple images for each person appearance. Therefore, ReID in multi-shot scenario is more suitable for practical applications. ReID in multi-shot scenarios is also referred to as video-based ReID.

[154] shows that temporal features (e.g. gait pattern) can offer discriminative features for person identification even using low resolution video sequences. In ReID multi-shot scenarios, these temporal features can be used in combination with spatial features to create better feature representation. Temporal features can improve the accuracy of ReID methods in particular when the majority of clothing worn by subjects tends to be non-discriminative [36, 124].

In [124], a deep learning video-based ReID method using a recurrent convolutional neural network architecture is proposed to exploit both spatial and temporal features. A single network is used to learn a representation for both feature types. This poses a limitation which constrains the amount of information that the network can learn. To address this limitation, we propose the use of a two stream convolutional neural network (CNN) [26] with weighted objective function where each stream has Siamese structure [155].

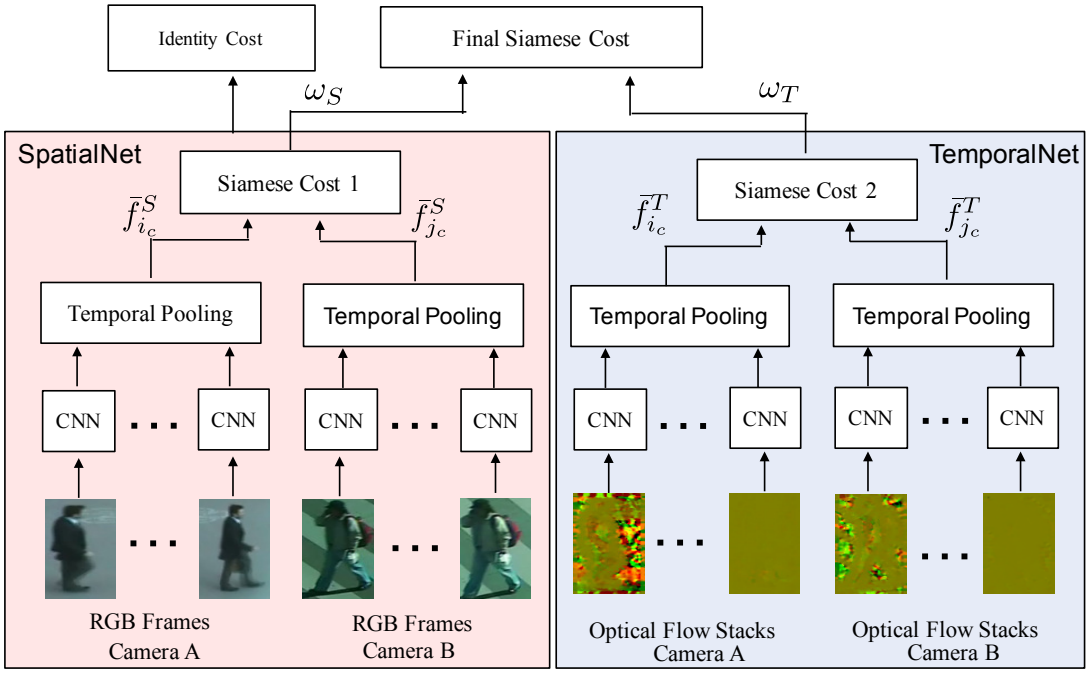


Fig. 4.1.: Overall Architecture of the proposed two stream ReID system

4.1 Proposed Method

4.1.1 Overall Network Architecture

The overall architecture of our proposed method is shown in Figure 4.1. The method is motivated by the fact that both spatial and temporal features possess discriminative information useful for the ReID task. However, the best feature representation does not need to be the same for both types of features. Therefore, we propose a two stream Siamese CNN which processes spatial and temporal information separately.

Siamese CNNs contain two identical sub-networks with shared weights and are suitable for tasks which involve finding the similarity between two comparable inputs [155]. CNNs typically process an image or multiple images and classify them into a single class, whereas Siamese CNNs process two images or two sequences of images and compute the similarity between them. In our proposed ReID system, the

input to first stream are two sequences of RGB frames where each sequence is captured from a different camera. The second stream processes the optical flow information from both cameras as shown in Figure 4.1. The input is described in more details in Section 4.1.2. Each stream is based on the same network architecture. Throughout this thesis, we will refer to the network associated with spatial content as SpatialNet and the network associated with temporal content as TemporalNet.

Both networks are composed of multiple CNNs with Siamese architecture, and all the CNNs within the same stream share the same parameters. We refer to this CNN as the "base CNN" and describe its structure in Section 4.1.3. The outputs of the base CNNs which processes images from the same camera view are combined using temporal pooling. The temporal pooling is described in Section 4.1.4. The outputs of the temporal pooling from both cameras are combined using the Siamese cost as described in Section 4.1.5. Finally, the two networks associated with both streams are fused together using a weighted cost function as described in Section 4.1.6.

4.1.2 The Inputs

We define the generic input sequence as I_c, where $c \in a, b$ for camera A and B, respectively. For the SpatialNet, the input sequence are RGB frames:

$$I_c = (S^{(1)}, \ldots, S^{(t)}, \ldots, S^{(L)}) \tag{4.1}$$

where L is the sequence length and $S^{(t)}$ is the RGB frame at time t.

For TemporalNet, optical flow images are used as input:

$$I_c = (T^{(1)}, \ldots, T^{(t)}, \ldots, T^{(L)}) \tag{4.2}$$

where L is the sequence length and $T^{(t)}$ is the input optical flow image at time t. The effectiveness of using optical flow to learn temporal features are demonstrated in [26, 124]. The optical flow is the pattern of apparent motion between two consecutive frames caused by the object or camera movement. It is the 2 dimensional vector where each vector is a displacement vector for horizontal and vertical direction. And this

2D vector is describing the movement of points from first frame to second frame. To define the problem, we assume that the motion will be small between two consecutive frames. The small motion in the images can be defined as

$$H(x, y) = I(x + u, y + v) \tag{4.3}$$

where $I(x, y)$ is the first frame, $H(x, y)$ is the second frame, u, v are the displacement of the object. We take the first order Taylor series expansion, then we can define $I(x + u, y + v)$ as

$$\begin{aligned} I(x + u, y + v) &= I(x, y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \text{higher order terms} \\ &\approx I(x, y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v. \end{aligned} \tag{4.4}$$

By combining Equations 4.3 and 4.4, we obtain

$$\begin{aligned} 0 &= I(x + u, y + v) - H(x, y) \\ &\approx I(x, y) + I_x u + I_y v - H(x, y) \\ &\approx (I(x, y) - H(x, y)) + I_x u + I_y v \\ &\approx I_t + \nabla I \cdot [u \quad v] \end{aligned} \tag{4.5}$$

With the limit $u \to 0, v \to 0$, Equation 4.5 becomes

$$I_t + \nabla I \cdot [\frac{\partial x}{\partial t} \quad \frac{\partial y}{\partial t}] = 0. \tag{4.6}$$

To solve Equation 4.6, we use the Lucas-Kanade optical flow technique [156]. In Lucas-Kanade technique, we assume that neighbor pixels within a small window have same displacements u and v. With this assumption, for each neighbor pixel p_i, we obtain the Equation 4.7 from Equation 4.6.

$$I_t(p_i) + \nabla I(p_i) \cdot [u \quad v] = 0 \tag{4.7}$$

We can form a matrix form equation by combining n equations from nXn window as

$$\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \vdots & \vdots \\ I_x(p_n^2) & I_y(p_n^2) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(p_1) \\ I_t(p_2) \\ \vdots \\ I_t(p_n^2) \end{bmatrix}. \tag{4.8}$$

This system has more equations than unknowns. Thus, Lucas-Kanade method obtains a solution by the least squares principle. Finally, the solution can be obtained as

$$d = (A^T A)^{-1} A^T b. \tag{4.9}$$

The advantage of LucasKanade method is that it is less sensitive to image noise than the point-wise methods due to the windowing process. However, since it is a local method, it can not provide the information in the entire image region.

To obtain optical flow image $T^{(t)}$ from displacement vectors, we add 127 to each value and divide by 255. Sample optical flow images are shown in Figure 4.1.

4.1.3 The Base CNN Architecture

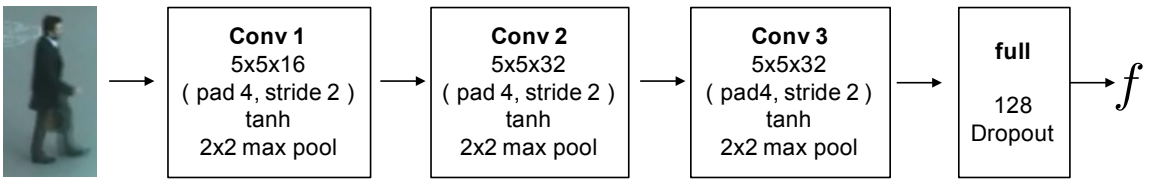


Fig. 4.2.: The structure of the base CNN and hyper-parameters

As shown in Figure 4.1, the input sequence I_c is processed using the base CNN. Figure 4.2 shows the base CNN structure and the hyper-parameters associated with it. The CNN takes one input sample ($S^{(t)}$ or $T^{(t)}$) and produces the output feature vector f^S or f^T for SpatialNet and TemporalNet, respectively. Our base CNN is composed of three convolution layers where each layer has convolution, non-linear activation and max-pooling steps. We use hyperbolic-tangent (tanh) as non-linear activation function. At the end of the three convolution layers, a fully connected layer is placed to have mapping to all the activations from the last convolution layer. Dropout [157] is also used to reduce the model over-fitting.

4.1.4 Temporal Pooling

For the SpatialNet, the base CNN processes a single RGB frame out of the sequence of frames in the multi-shot scenario, or optical flow content in the case of the TemporalNet. Combining the spatial or temporal features using multiple frames helps address the challenges associated with various viewpoints and poses. To process the input sequence, each sub-network of the Siamese network in each stream utilizes L base CNNs and produces L feature vectors. The feature vectors produced by the L CNNs in each sub-network are combined into a single feature vector using temporal pooling. Max pooling, sum pooling and mean pooling are the most common techniques used to achieve this. In [124], the RNN-ReID method has shown that the mean pooling method is the most suitable temporal pooling technique for the ReID task. We adopt the same approach in our proposed method. If we denote the base CNN by the function $C()$, then the temporally pooled feature vector, $\bar{f}_{i_c}$, is computed as follows:

$$\bar{f}_{i_c} = \frac{1}{L} \sum_{t=1}^{L} C(I_{i_c}^{(t)}) \tag{4.10}$$

where i is the person ID, $c \in \{a, b\}$ is the camera view and $I_{i_c}^{(t)}$, $t \in 1, \ldots, L$, is one element (RGB image or optical flow vectors) of the input multi-shot sequence. The sequence of images are processed and temporally pooled to obtain the feature vector $\bar{f}_{i_c}^{S}$ or $\bar{f}_{i_c}^{T}$ for the SpatialNet and TemporalNet, respectively.

4.1.5 Siamese Cost

Siamese networks are composed of two sub-networks with shared weights [155]. While learning the features from each sub-network, Siamese networks compare the features from the pair using Euclidean distance. Thus, in training process, the network tries to minimize the distance between feature pairs when they are from the same class and maximize the distance between feature pairs when they are from different classes. Due to this property, Siamese networks have been widely used for the ReID

task since the goal is to find the similarity between a pair of sequences. As mentioned before, we use a Siamese network for both streams: SpatialNet and TemporalNet as shown in Figure 4.1. Furthermore, the generic Siamese cost of our proposed method can be defined as follows:

$$D(\bar{f}_{i_c}, \bar{f}_{j_c}) = \begin{cases} \frac{1}{2}||\bar{f}_{i_c} - \bar{f}_{j_c}||^2, & \text{if } i = j \\ \frac{1}{2}\{\max(m - ||\bar{f}_{i_c} - \bar{f}_{j_c}||, 0)\}^2, & \text{if } i \neq j \end{cases} \tag{4.11}$$

where m is the Siamese margin and $\bar{f}_{i_c}$, $\bar{f}_{j_c}$ are the temporally pooled feature vectors for person i and j, respectively. Equation 4.11 applies to both SpatialNet and TemporalNet in the same way with different type of inputs.

4.1.6 Weighted Two Stream Joint Identification and Verification

During the training process, we build on the joint identification and verification approach from [158] to define our training objective. We use the softmax loss function to compute the identification cost as in [124]. Then, this cost is integrated into our final training objective function as explained later. The identification cost is defined as:

$$V(x) = P(q = c|x) = \frac{\exp(W_c x)}{\sum_k \exp(W_k x)} \tag{4.12}$$

where x is the feature vector and q is the person's identity. W_c and W_k indicate the cth and the kth column of the softmax matrix W, respectively. Note that the softmax matrix W is the matrix representation of the fully connected layer in the base CNN architecture.

From the RNN-ReID method, it was already observed that joining the identification with the Siamese cost is crucial to improve the ReID accuracy. We have two

Siamese cost functions from each stream, whereas RNN-ReID has only one Siamese cost. Therefore, we define the combined cost function J_f as follows:

$$\begin{aligned} J_f =& \omega_S D(\bar{f}^S_{i_c}, \bar{f}^S_{j_c}) + \omega_T D(\bar{f}^T_{i_c}, \bar{f}^T_{j_c}) \\ &+ V(\bar{f}^S_{i_c}) + V(\bar{f}^S_{j_c}) \end{aligned} \tag{4.13}$$

where V is the standard softmax loss defined in Equation 4.12. ω_S, ω_T are the weights for SpatialNet and TemporalNet, respectively. Note that we only use the identification cost V which is computed using the spatial features since they contain more information regarding to the person label than the temporal features. We propose using different weights for each stream to be able to emphasize the spatial features as compared to the temporal features. For ReID Task, even though walking motion adds discriminative power to the ReID solution, spatial features such as appearance, color or texture are relatively more important in terms of re-identifying people. Thus, we set the weights empirically with the condition $\omega_S \geq \omega_T$.

4.1.7 Similarity Metric for Testing

The weighted two stream joint identification and verification objective function, which is used for training, incorporates the ability to predict a person's identity. However, during the evaluation, the goal is to find the similarity score (metric) between two sequences of images and to rank the gallery accordingly. Therefore, we modify Equation 4.13 to disregard the contribution of the standard softmax loss V and replace the Siamese cost D with the Euclidean distance. The Euclidean distances are computed using the temporally pooled feature vectors ($\bar{f}^S_{i_c}$, $\bar{f}^T_{i_c}$, $\bar{f}^S_{j_c}$ and $\bar{f}^T_{j_c}$) as follows:

$$d_S = ||\bar{f}^S_{i_a} - \bar{f}^S_{j_b}|| \tag{4.14}$$

$$d_T = ||\bar{f}^T_{i_a} - \bar{f}^T_{j_b}|| \tag{4.15}$$

Finally, d_S and d_T are combined using a weighted average to compute the final similarity metric d_F:

$$d_F = \frac{\omega_S d_S + \omega_T d_T}{\omega_S + \omega_T} \tag{4.16}$$

4.2 Experiment Results

In this section, we evaluate our proposed method using the publicly available datasets: Person re-identification (PRID2011) dataset [35] and the iLIDS video re-identification (iLIDS-VID) dataset [36]. We investigate our proposed method with different hyper-parameter settings and evaluate the performance against the state-of-the-art ReID methods.

4.2.1 Datasets

Both datasets feature a multi-shot scenario in which a person trajectory is represented by a sequence of images. The PRID2011 dataset contains images from two non-overlapping static surveillance cameras. The sequence presents the significant differences in viewpoint, illumination and camera characteristics. It is composed of 385 person trajectories from one view and 749 from the other one, with 200 persons appearing in both views. Each image sequence has a variable length ranging from 5 to 675 image frames, with an average number of 100 images. We only consider the 200 persons appearing in both views as suggested in [36].

The iLIDS-VID dataset was created by observing pedestrians in two camera views. The outputs of two non-overlapping cameras were captured at a crowded airport arrival hall. It consists of 600 image sequences of 300 individuals with one pair of sequences from two camera views for each person. Each image sequence has a variable length ranging from 23 to 192 image frames, with an average number of 73 images. It is one of the most challenging datasets due to the cluttered background and random occlusions.

4.2.2 Experiment Setup

Input images are pre-processed before being fed into the two stream Siamese CNN. Each color channel of the RGB image is normalized to introduce invariance

to illumination changes. This is simply done by subtracting the mean and dividing by the standard deviation. Each horizontal and vertical optical flow channel is also normalized to the range of $[-1, 1]$.

The same data augmentation technique in [124] is used to add more variety to the data. Random mirroring and cropping are used for data augmentation. Note that a consistent data augmentation technique is applied to the images from the same sequence.

As suggested in [124], positive and negative pairs are alternatively fed into the network. Sequence pairs are randomly sampled from the all training identities. All training sequence lengths are set to 16 and the test sequence lengths are varied to investigate the significance of the sequence length as described in Section 4.2.6. Note that this sequence length can be arbitrary due to the network architecture.

The proposed network is trained for 1000 epochs using the stochastic gradient descent method. The batch size is set to 1, the learning rate to $1e^{-3}$ and the momentum to 0.9. The Siamese cost function margin is set to $m = 2$. The base CNN feature dimension is 128 with the dropout rate set to 0.5.

4.2.3 Evaluation Protocol

We follow the evaluation protocol described in [36]. The dataset is randomly split into two subsets with the same size. One is used for training and one for testing. For the testing, the sequences from the first camera are used as the probes while the sequences from the second camera are used as the gallery.

We validate the performance of our proposed method and compare the performance against other methods using the Cumulative Matching Characteristic (CMC) curve which indicates the probability of finding the correct match in the top K matches within the ranked gallery. The experiment is repeated five times by randomly splitting the dataset into training and testing and the average result is reported.

In our proposed method, we have two extra hyper-parameters (ω_S, ω_T). To see the effectiveness of proposed method, we perform experiments with various hyper-parameters settings. We perform experiments with $\omega_S = 1$ when ω_T is set to 0 or 1 in order to verify the individual contribution of TemporalNet. We also perform experiments with $\omega_S = 2, 3$ when $\omega_T = 1$ to see the relative contribution of the spatial features as compared to the temporal features.

4.2.4 Training Details

In training process, we proposed a new cost function (J_f) for proposed network as described in Equation 4.13. We defined the cost computed with the given input at specific epoch as Loss. Loss changes throughout each epoch are provided in Figure 4.3 and Figure 4.4. Since our cost function has new hyper-parameter ω_S, we plot the loss value with $\omega_S = 1$, 2, 3 which are same as experiment setting. As increasing the ω_S, the total loss value increases since J_f is proportional to ω_S. Note that loss value from ILIDS-VID [36] dataset is much higher than the one from PRID2011 [35]. It is expected behavior since ILIDS-VID dataset has more challenges such as cluttered background or occlusions.

4.2.5 Feature Visualization

One of the most common and widely used deep learning feature visualization method is visualizing the convolution layer output given the input [107]. We use this technique to visualize proposed network features. Proposed method has three convolution layers in base CNN structure as shown in Figure 4.2.

Note that conv1 layer has 16 filters and conv2, conv3 layers have 32 filters. A sample is randomly selected from each dataset (PRID2011 [35] and ILIDS-VID [36] dataset). This sample passes through the forward pass with the trained proposed network to visualize the features in the middle convolution layers (conv1, conv2 and conv3 in Figure 4.2).

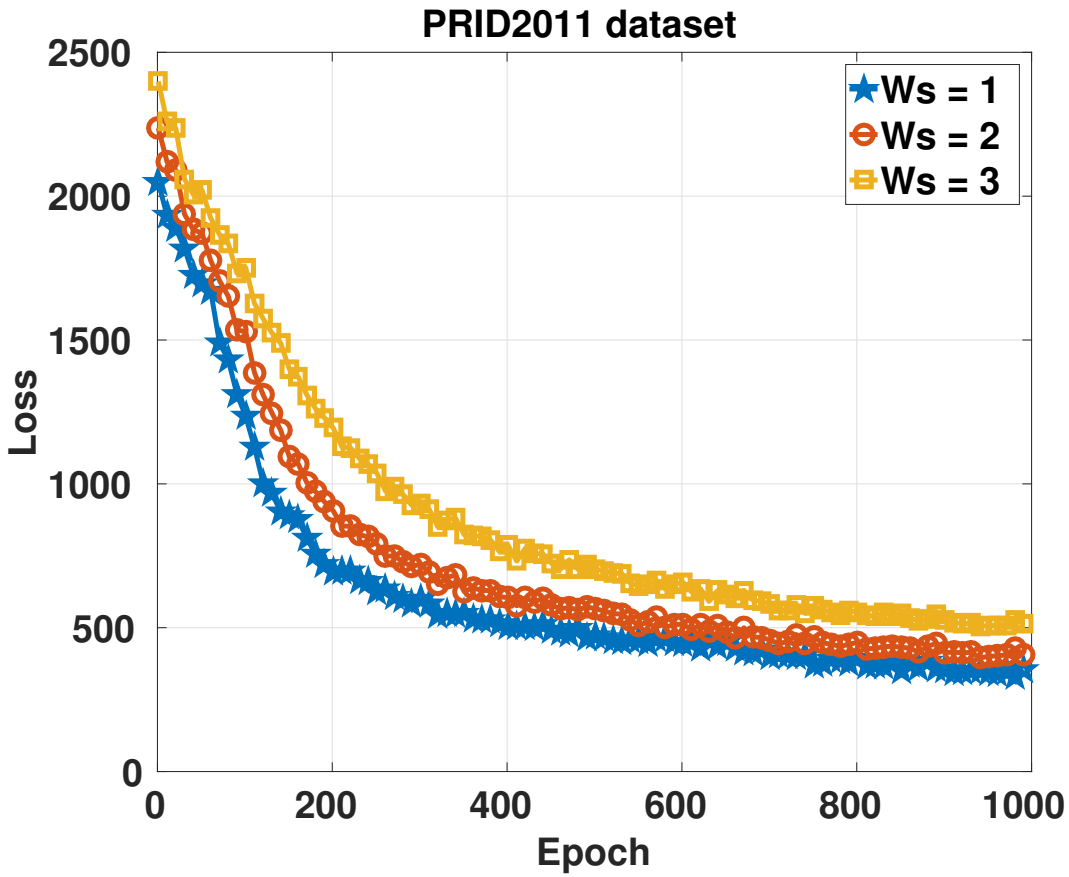


Fig. 4.3.: Loss function over epochs in PRID2011 [35]

For the SpatialNet, Figure 4.5, Figure 4.6, Figure 4.9 and Figure 4.10 show the features associating with the RGB image as input. As we expected, the features learned from the SpatialNet show mostly appearance information such as edges and objects in the image. These features are critical for ReID Task. For the TemporalNet, Figure 4.7, Figure 4.8, Figure 4.11 and Figure 4.12 show the features associating with the optical flow image as input. Since the optical flow image had less structural information from the start, it shows less structural information in the visualized feature images.

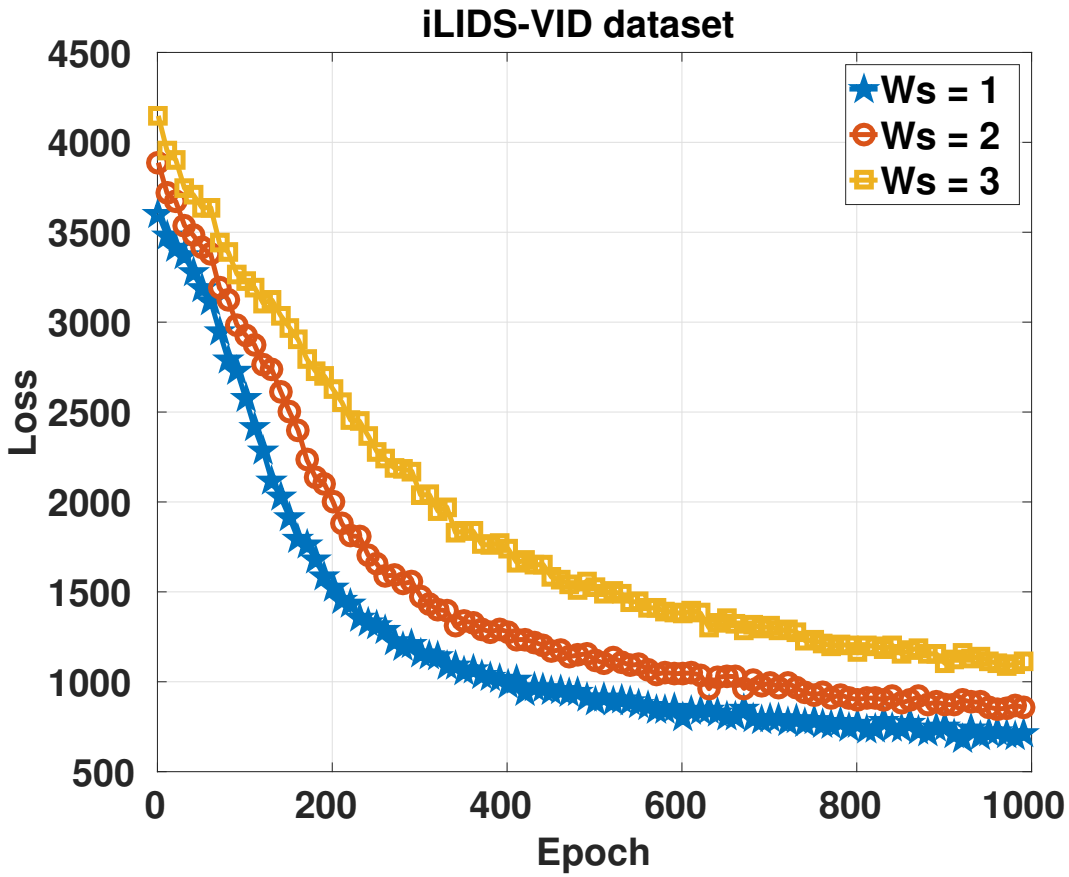


Fig. 4.4.: Loss function over epochs in ILIDS-VID [36]

Fig. 4.5.: Visualized features of SpatialNet with the sample from camera A in PRID2011 [35]

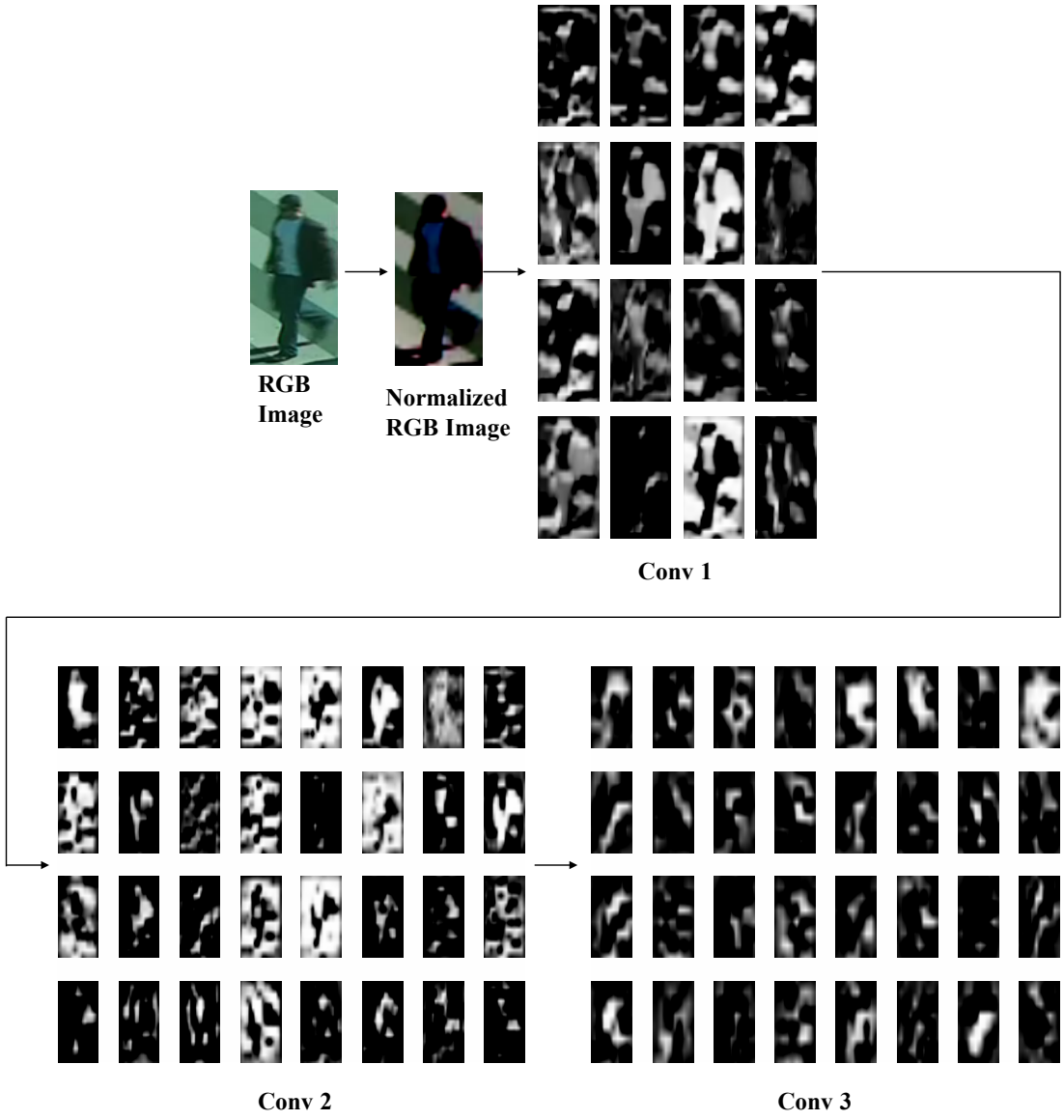


Fig. 4.6.: Visualized features of SpatialNet with the sample from camera B in PRID2011 [35]

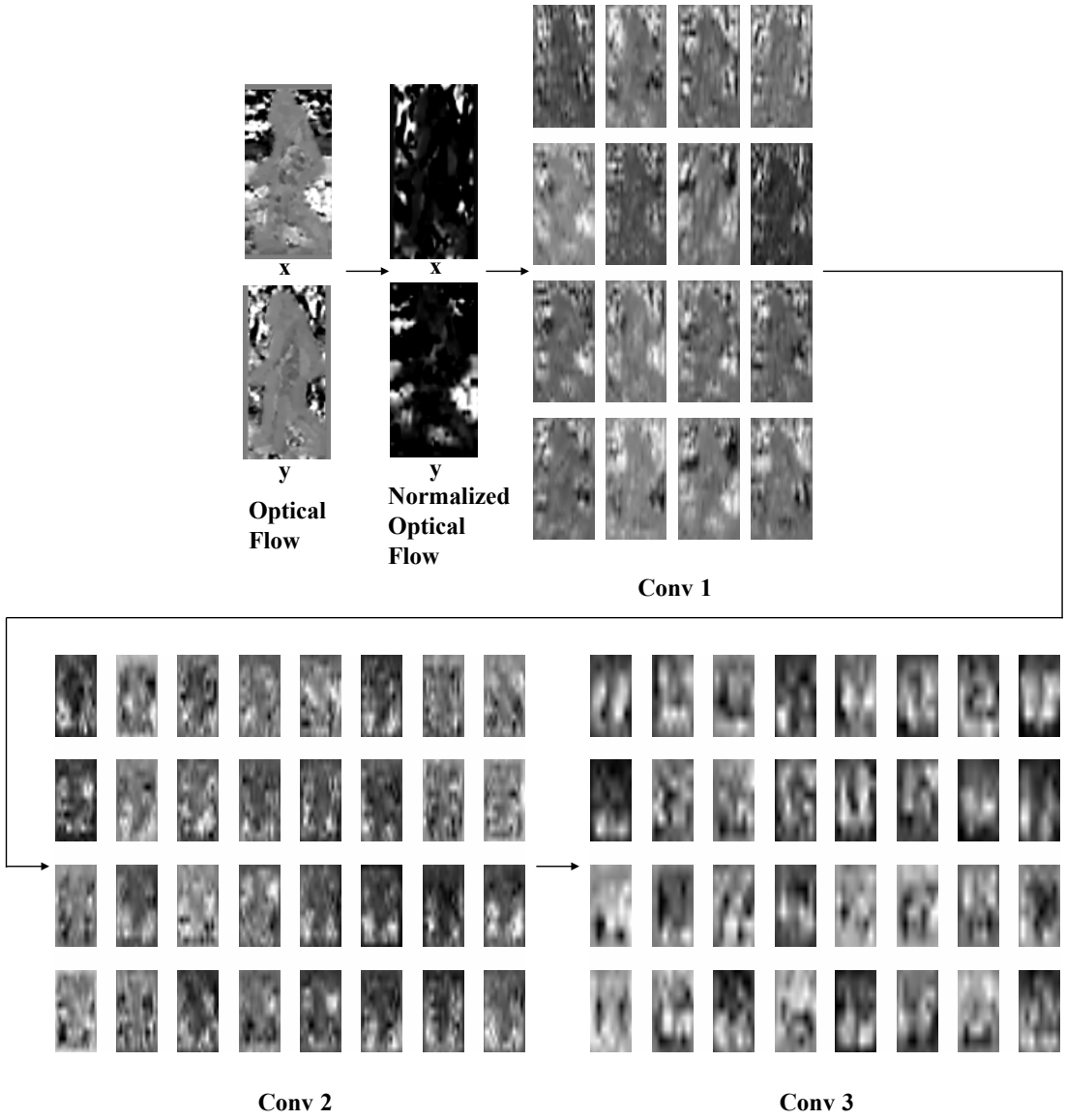


Fig. 4.7.: Visualized features of TemporalNet with the sample from camera A in PRID2011 [35]

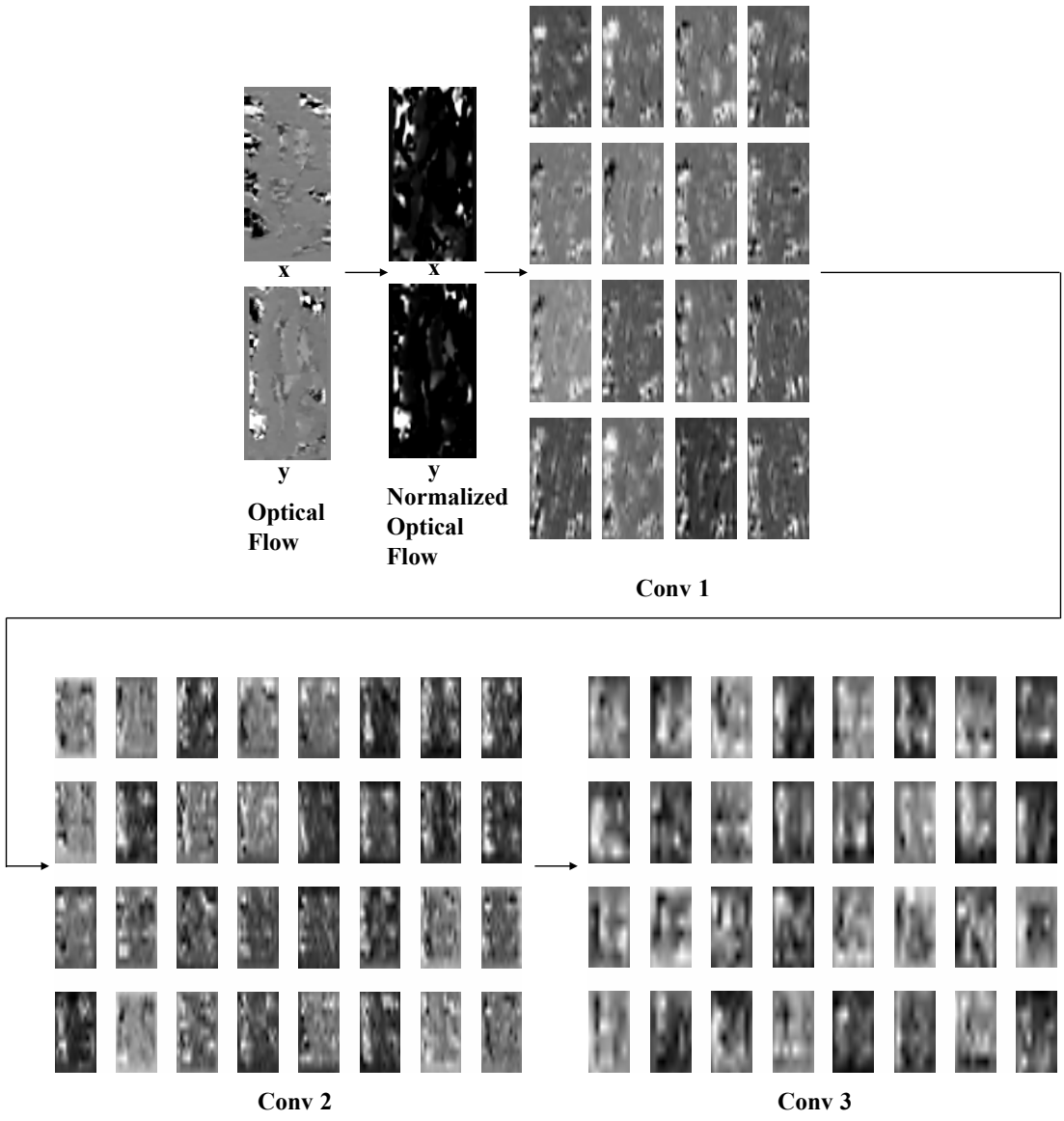


Fig. 4.8.: Visualized features of TemporalNet with the sample from camera B in PRID2011 [35]

Fig. 4.9.: Visualized features of SpatialNet with the sample from camera A in ILIDS-VID [36]

Fig. 4.10.: Visualized features of SpatialNet with the sample from camera B in ILIDS-VID [36]

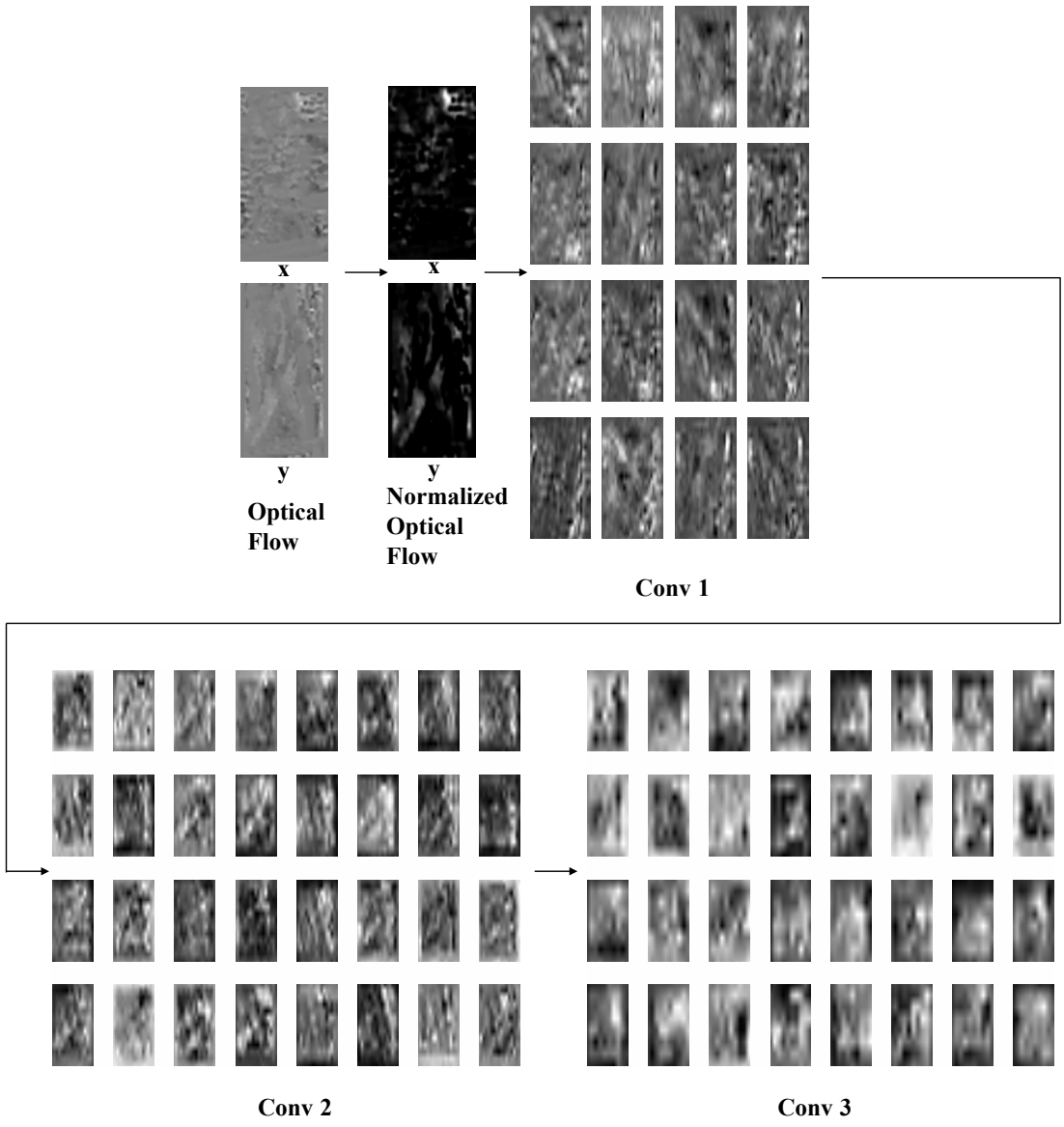


Fig. 4.11.: Visualized features of TemporalNet with the sample from camera A in ILIDS-VID [36]

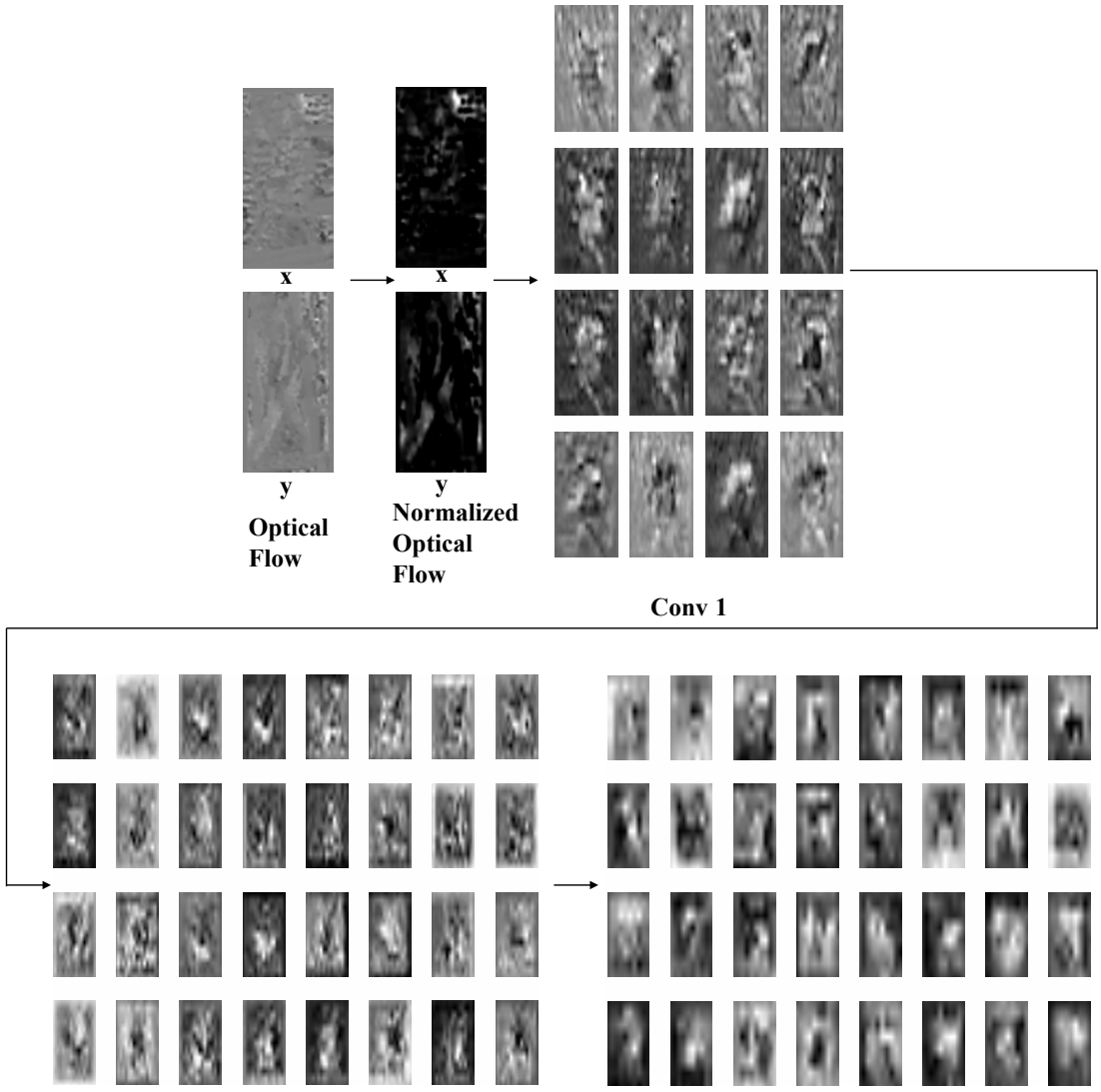


Fig. 4.12.: Visualized features of TemporalNet with the sample from camera B in ILIDS-VID [36]

4.2.6 Results and Discussion

Probe and Gallery Sequence Length

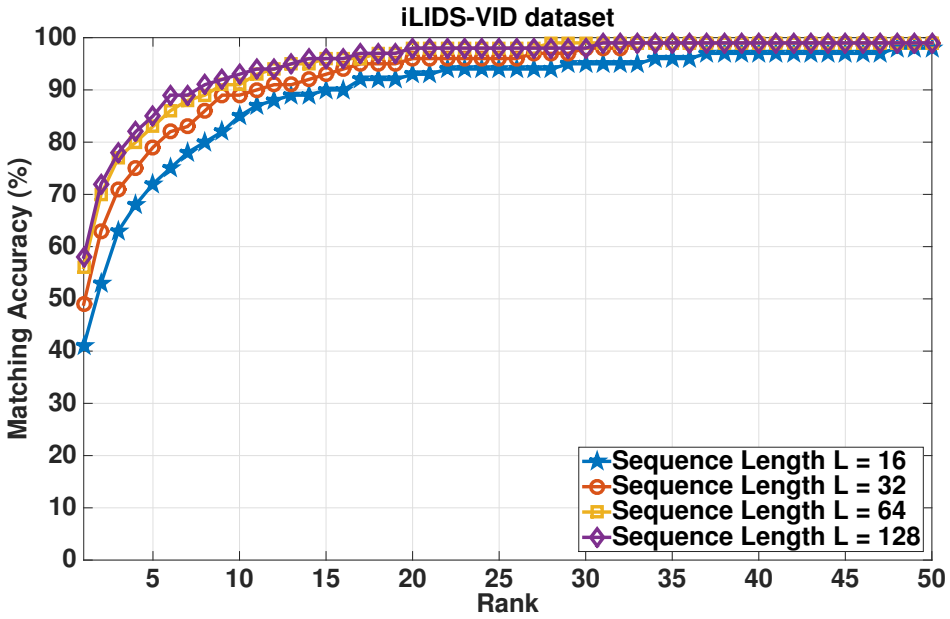


Fig. 4.13.: CMC curves for different probe/gallery sequence lengths

Table 4.1.: Matching accuracies with various probe/gallery sequence lengths in iLIDS-VID

Length \ Rank	1	5	10	20
16	41	70	81	92
32	50	79	88	95
64	56	82	91	97
128	58	85	93	97

In this section, we investigate the significance of the sequence length during testing. An experiment is conducted to evaluate the ReID matching accuracy using various sequence lengths. Our proposed network shown in Figure 4.1 is trained with the sequence length set to 16 using the iLIDS-VID dataset. During evaluation, the matching accuracy is calculated using $\{16, 32, 64, 128\}$ as lengths for the probe and gallery sequences. In the case when the probe or gallery sequence is shorter than the test length, we use the entire sequence.

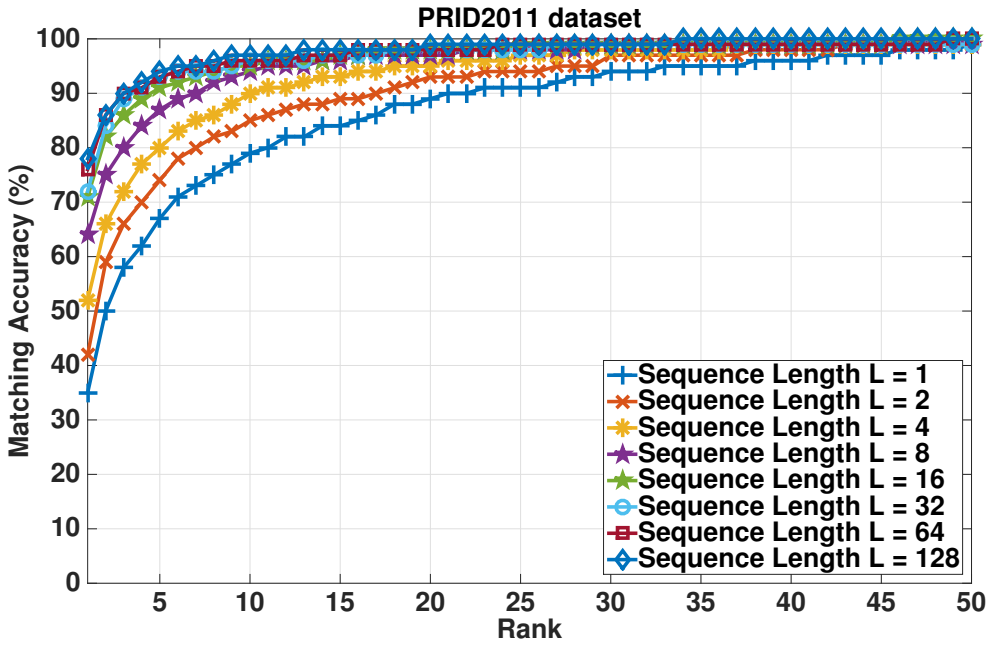


Fig. 4.14.: Matching Accuracy for variable sequence Length in PRID2011 [35]

The matching accuracies for different sequence lengths are summarized in Table 4.1. The results clearly indicate that the matching accuracies are improved as the sequence length is increased. For instance, when we increase the sequence length from 16 to 128, the top rank matching accuracy is improved by 17%. This is an intuitive result since combining the spatial and temporal features using multiple images helps address the challenges associated with various viewpoints and poses.

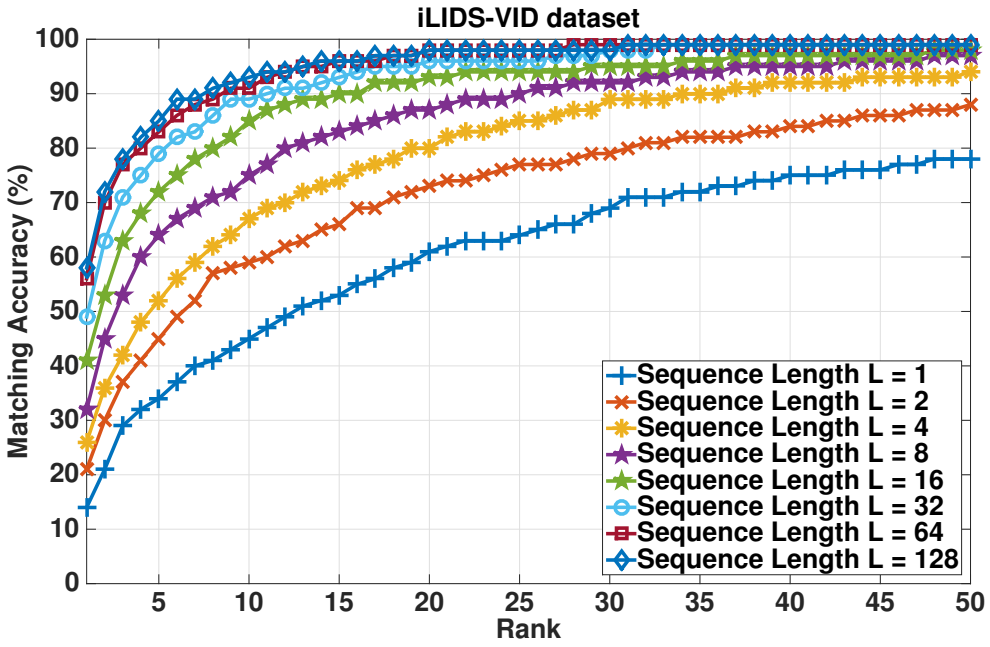


Fig. 4.15.: Matching Accuracy for variable sequence Length in ILIDS-VID [36]

Figure 4.14 and Figure 4.15 show extra experiment results from the shorter sequence length 1 to 128. Compare to the case using only single image, the matching accuracy has increased significantly when we use 128 sequence length for both datasets.

Verification on Two Stream

To verify the usefulness of temporal information in ReID task, we perform the experiments with the different settings of the hyper-parameters (ω_S, ω_T). This also can verify the improvement gained by the use of a two stream CNN architecture. Note that ω_S and ω_T control the individual contributions of the SpatialNet and the TemporalNet, respectively. When $\omega_T = 0$, the contribution of TemporalNet becomes totally 0 in training phase. This also applies to the test phase in the same way based on the Equation 4.16.

Table 4.2.: Matching accuracies with different stream settings

(a) PRID2011

Streams \ Rank	1	5	10	20
$\omega_S = 1$, $\omega_T = 0$	75	93	97	98
$\omega_S = 1$, $\omega_T = 1$	78	94	94	99
$\omega_S = 2$, $\omega_T = 1$	78	94	97	99
$\omega_S = 3$, $\omega_T = 1$	79	93	97	98

(b) iLIDS-VID

Streams \ Rank	1	5	10	20
$\omega_S = 1$, $\omega_T = 0$	57	60	91	95
$\omega_S = 1$, $\omega_T = 1$	58	86	93	97
$\omega_S = 2$, $\omega_T = 1$	60	86	93	97
$\omega_S = 3$, $\omega_T = 1$	56	86	92	96

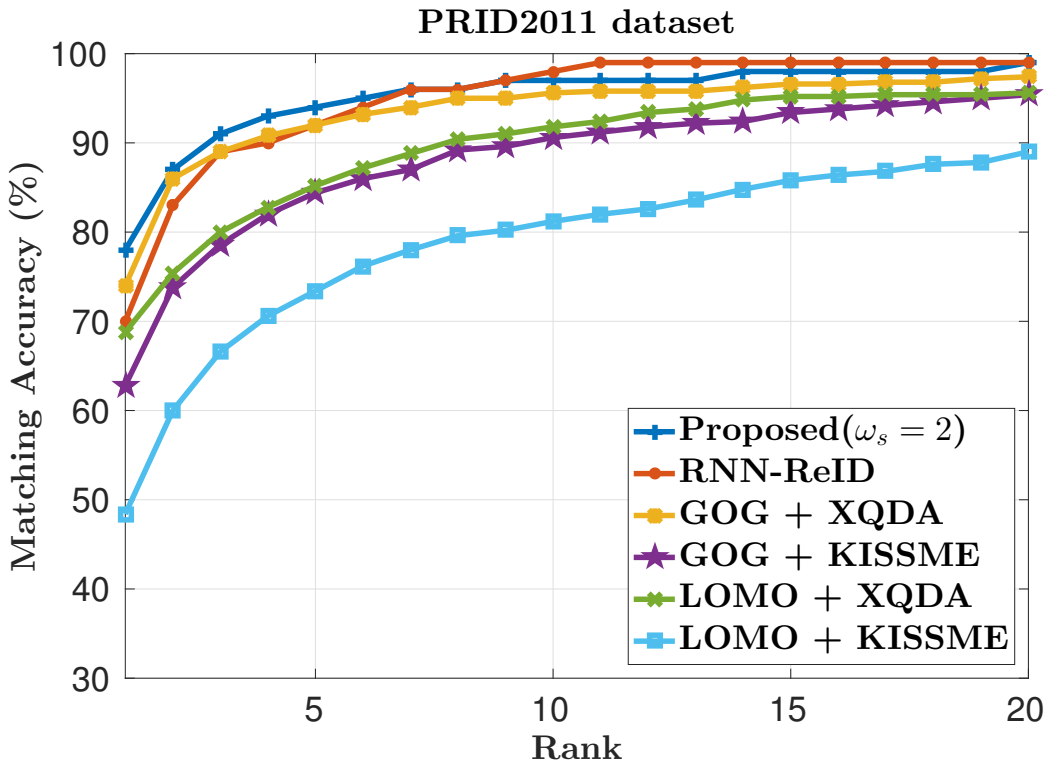
We then compare ReID matching accuracies for different hyper-parameter settings such as spatial only case ($\omega_T = 0$, $\omega_S = 1$,) and Both Stream cases when ω_T is fixed to 1 while ω_S is varying from $1-3$. As shown in Table 4.2, using both stream cases have 3-4% accuracy improvement in PRID2011 and 1-3% accuracy improvement in iLIDS-VID. This result demonstrates that by having two separate networks to represent the spatial and the temporal content, each network is able to learn the best feature representation and improves the ReID performance. In addition, based on the results for $\omega_S \geq 2$ cases, ReID performance improved in PRID2011 whereas it did not improve in iLIDS-VID for $\omega_S = 3$ case. We thus conclude that the optimal relative contribution of the spatial and temporal features is data dependent.

Comparisons

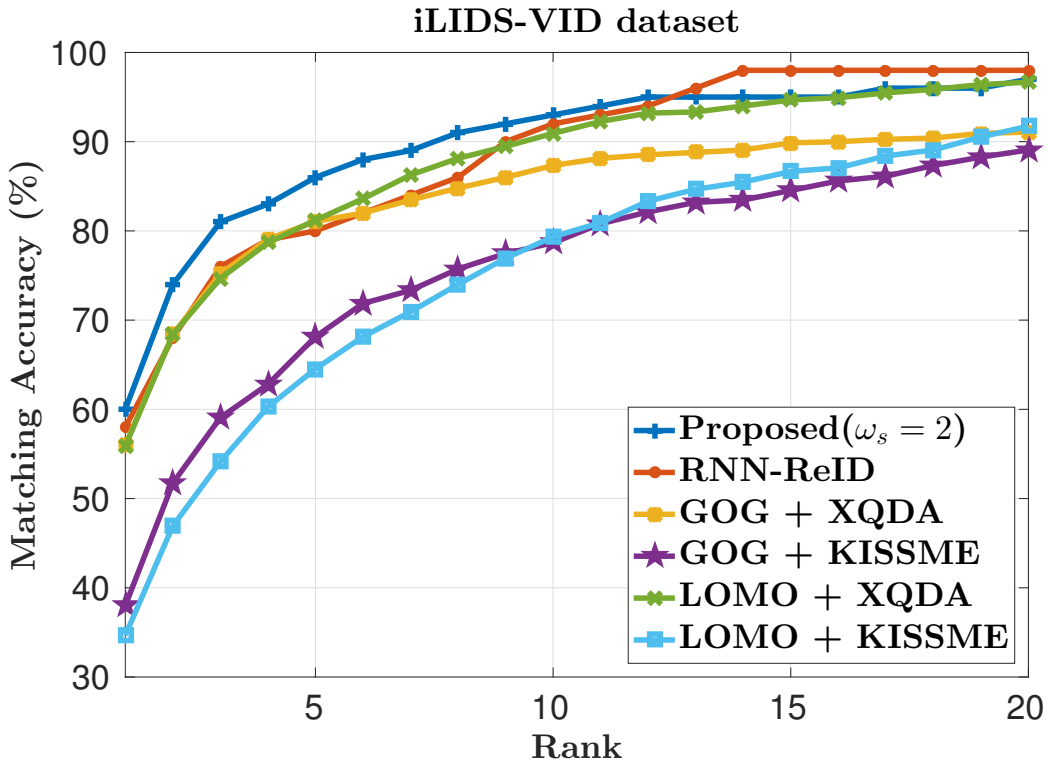
We compare the performance of our proposed method against several of the best performing methods in a multi-shot ReID setting. We evaluate state-of-the-art metric learning methods (XQDA [41] and KISSME [68]) using state-of-the-art feature extraction methods: LOMO [41], GOG [67] and ELF [66]. Since we are evaluating multi-shot ReID methods, we extract the features for each image in the sequence and compute the average which is used by the metric learning methods. To our best knowledge, the combination of GOG and XQDA achieves state-of-the-art performance and the RNN-ReID method is the best performing deep learning method [124].

The CMC curves are plotted in Figure 4.16a and 4.16b and the matching accuracies are summarized in Table 4.3a and 4.3b for the PRID2011 and the iLIDS-VID datasets, respectively. For the PRID2011 dataset, our proposed method outperforms all the other methods. The top rank matching accuracy is 4% higher than the accuracy achieved by the GOG+XQDA method and 8% higher than RNN-ReID.

For the iLIDS-VID dataset, the results show that our approach has comparable accuracy to the RNN-ReID method and is 5% higher than the accuracy achieved by the GOG+XQDA method as can be seen in Table 4.3b and Figure 4.16b. The top rank matching accuracy for the iLIDS-VID dataset is 18% lower than the case for the PRID2011 dataset. We believe this mainly due to the cluttered background and occlusions associated with the iLIDS-VID dataset.



(a) PRID 2011.



(b) iLIDS-VID.

Fig. 4.16.: CMC Curves for comparison.

Table 4.3.: Matching accuracies comparison with previous methods

(a) PRID2011

Methods \ Rank	1	5	10	20
Proposed ($\omega_S = 2$)	78	94	97	99
RNN-ReID [124]	70	92	98	99
GOG [67] + XQDA [41]	74	91	94	96
GOG [67] + KISSME [68]	57	80	89	94
LOMO [41] + XQDA [41]	67	86	92	94
LOMO [41] + KISSME [68]	48	72	82	91
ELF [66] + XQDA [41]	22	43	54	64
ELF [66] + KISSME [68]	15	32	42	56

(b) iLiDS-VID

Methods \ Rank	1	5	10	20
Proposed ($\omega_S = 2$)	60	86	93	97
RNN-ReID [124]	58	80	92	98
GOG [67] + XQDA [41]	55	79	86	90
GOG [67] + KISSME [68]	38	67	79	89
LOMO [41] + XQDA [41]	53	79	88	95
LOMO [41] + KISSME [68]	35	65	79	90
ELF [66] + XQDA [41]	23	49	60	74
ELF [66] + KISSME [68]	15	40	55	70

5. SIMILARITY PRESERVING STARGAN FOR PERSON RE-IDENTIFICATION

As deep learning approaches have been studied, large-scale ReID datasets have been released : Market-1501 [38] and DukeMTMC-reID [37]. Compare to the other datasets such as PRID 2011 [35], iLIDS-VID [36], Market-1501 and DukeMTMC-reID have more than 6 different cameras settings and a large number of images and identities. This means that the same person can show up in more than two camera views which introduces more challenges to re-identify the person.

Although larger datasets have been introduced, more training data is needed. In addition, since the number of camera is growing in these datasets, more samples for each camera are needed in order to learn robust camera invariant feature representations. It is expensive and time-consuming to have manual identity annotations across different cameras as we have more cameras and videos to annotate the identity for ReID tasks.

To alleviate this problem, Zhong *et al.* [43] proposed a method for generating scene style transferred images using a CycleGAN (CamStyle) [133] as a data augmentation method for ReID. They trained multiple image-to-image translation models for each camera pair using CycleGAN. Then, the model can generate new sample images from the source style to the target style. In CamStyle [133], authors defined this problem as generating "Camera Style" transferred images. However, in this problem, we do not consider camera specific settings such as focal length or angle of view of the camera. Thus, we re-define this problem as "Scene Style" transfer problem. Scene style means the scene specific characteristics across different cameras such as bright or dark illumination. We will refer generated images as scene style transferred images in the rest of this thesis. These scene style-transferred images allow us to have extra training samples with different scene styles without additional manual

annotation. In other words, we have extra images as if they are captured from target camera but it is actually generated from the real image captured from source camera. In addition, label smoothing regularization (LSR) on the style-transferred images to softly distribute their labels and reduce the noise effect generated by the extra samples. Due to the limitation of CycleGAN which can model only one-to-one domain mapping, this method only can learn the mapping for one camera pair (e.g., camera 1 to camera 2) with the single model. Thus, using Camstlye [133], multiple models need to be trained to model an entire camera network. For example, in the DukeMTMC-reID [37] dataset which has 8 different cameras, $C_8^2 = 28$ different models need to be trained separately. The time complexity as well as the the number of parameters will be sharply increased as the number of camera increases. In addition, cross-camera relationships will be ignored in this architecture.

To address these limitations, we propose the use of StarGAN [42] with an additional similarity preserving term in the loss function for scene-aware image-to-image translation to generate the extra samples for ReID.

5.1 Proposed Method

Figure 5.1 shows the overall flow of our proposed method. First, we train the similarity preserving StarGAN to obtain the scene-aware image-to-image translation model. This model learns the mapping across different cameras with a single model in the ReID dataset. We then generate scene style translated images for all respective camera combinations from this single model. Finally, we train the deep learning ReID network with both real images and scene style translated images.

5.1.1 StarGAN

In this section we briefly revisit the StarGAN [42]. StarGAN has a single generator G learning the mappings among multiple domains and a single discriminator D with auxiliary classifier to discriminate fake and real images and control multiple domain

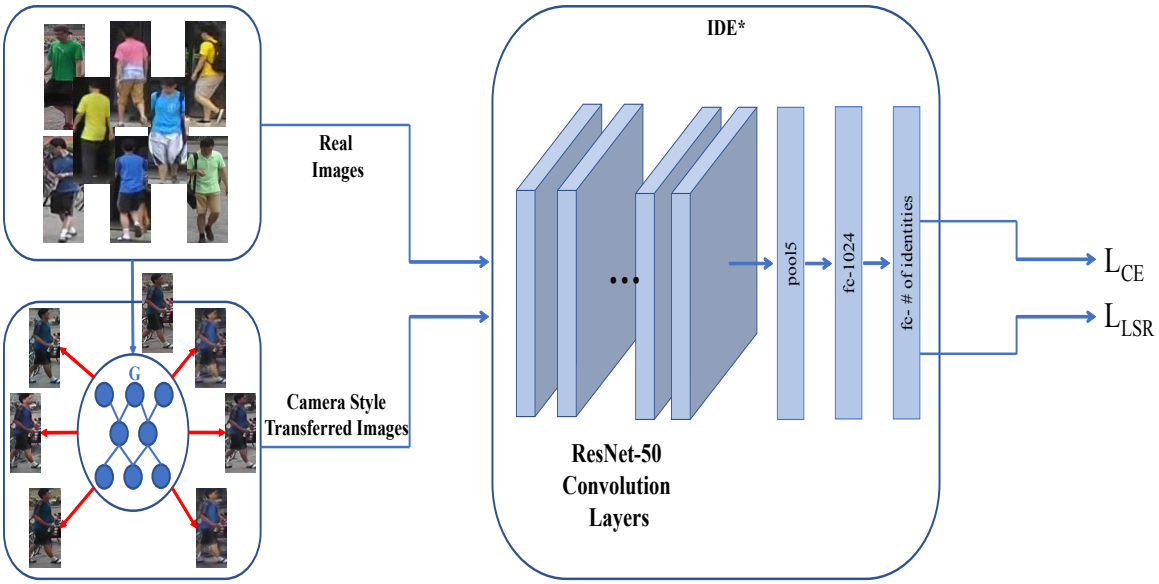


Fig. 5.1.: Overall Proposed Framework

simultaneously. In order to stabilize the training process while generating realistic fake images, the Wassertein GAN loss with a gradient penalty [159, 160] was used for the adversarial loss and defined as:

$$\begin{aligned} L_{adv} =& \mathbb{E}_x[D_s(x)] - \mathbb{E}_{x,c_t}[D_s(G(x,c_t))] \\ & - \lambda_{gp}\mathbb{E}_{\hat{x}}[(||\nabla_{\hat{x}} D_s(\hat{x})||_2 - 1)^2] \end{aligned} \tag{5.1}$$

where D_s is defined as the probability distribution over the sources, $\hat{x}$ is uniformly sampled along a straight line between a pair of a real and a generated image. In addition, G generates an image $G(x, c_t)$ mapped from the input image x to the target domain label c_t, while D tries to distinguish the between real and generated images.

StarGAN [42] has an auxiliary classifier on top of D to classify images to the respective domain label. For the real image, a domain classification loss is defined as:

$$L^r_{cls} = \mathbb{E}_{x,c_s}[-logD_{cls}(c_s|x)] \tag{5.2}$$

where $D_{cls}(c_s|x)$ denotes the probability distribution over domain labels given the real image x and c_s means the source domain label. For the fake image, a domain classification loss is described as:

$$L_{cls}^{f} = \mathbb{E}_{x,c_t}[-logD_{cls}(c_t|G(x,c_t))] \tag{5.3}$$

where $D_{cls}(c_t|G(x,c_t))$ represents a probability distribution over domain labels given the fake image $G(x,c_t)$ and c_t refers the target domain label.

In order to preserve the content of the input images while translating the domain-related information of the image, StarGAN used a cycle consistency loss [133] which is defined as

$$L_{rec} = \mathbb{E}_{x,c_t,c_s}[||x - G(G(x,c_t),c_s)||_1] \tag{5.4}$$

where the translated image $G(x,c_t)$ becomes the input for the G with the original domain label c_s and reconstruct the original image x.

Finally, the overall StarGAN loss function is expressed as

$$L_D = -L_{adv} + \lambda_{cls}L_{cls}^{r}, \tag{5.5}$$

$$L_G = L_{adv} + \lambda_{cls}L_{cls}^{f} + \lambda_{rec}L_{rec} \tag{5.6}$$

where λ_{cls}, λ_{rec} are hyper-parameters for the relative importance of each term.

5.1.2 Similarity Preserving StarGAN

In this thesis, we employ the StarGAN model to generate scene style translated images as extra training samples for ReID. However, we observe that the cycle consistency term, L_{rec} in Equation 5.4, was not enough for preserving the content of the input image related to person identity while translating the camera domain-related content. For scene-aware image-to-image translation, we do not want to have dramatic changes in the image since we need to keep the same identity while transferring the scene specific settings. In order to preserve the same identity while transferring the image to the different camera setting, we add two additional terms into StarGAN

generator loss (Equation 5.6). We present the details of each additional term in the following.

Identity Mapping Loss.

In order to preserve the cycle-consistency between the input and output, we add the identity mapping loss [161] to regularize the generator to be an identity mapping when the real image with the source domain label is provided.

The identity mapping loss term is defined as

$$L_{id} = \mathbb{E}_{x,c_s}[||G(x, c_s) - x||_1] \tag{5.7}$$

where $G(x, c_s)$ is the generated image with the source camera label c_s and the x is the real image from camera c_s.

Multi-scale Structural Similarity.

Wang *et al.* [162] originally used the structural similarity between two images across different scales. We add the multi-scale SSIM (MS-SSIM) term to our generator loss to preserve the structural similarity. Specifically in scene-aware image translation, we need to preserve the most of the structural information to maintain the same identity. By using this term, the generator tries to preserve the structural information of the input image.

Since SSIM is a single scale approach, the algorithm depends on the specific viewing conditions such as image resolution or viewing distance. To address this drawback, multi-scale SSIM (MS-SSIM) [162] was introduced to calibrate the parameters that control the relative importance across different scales. In ReID, we may have very different input image resolutions. Thus, we employ the MS-SSIM for the similarity metric in our proposed method.

Let $x_r = G(G(x, c_t), c_s)$ as the reconstructed image with the source camera label c_s, c_t refers to the target camera label and the x as the input image. The SSIM loss can defined as

$$L_{SSIM}(x_r, x) = [l(x_r, x)^{\alpha} c(x_r, x)^{\beta} s(x_r, x)^{\gamma}] \tag{5.8}$$

where

$$l(x_r, x) = \frac{2\mu_{x_r}\mu_x + C_1}{{\mu_{x_r}}^2 + {\mu_x}^2 + C_1} \tag{5.9}$$

$$c(x_r, x) = \frac{2\sigma_{x_r}\sigma_x + C_2}{{\sigma_{x_r}}^2 + {\sigma_x}^2 + C_2} \tag{5.10}$$

$$s(x_r, x) = \frac{\sigma_{x_r x} + C_3}{\sigma_{x_r}\sigma_x + C_3}. \tag{5.11}$$

$l(x_r, x)$, $c(x_r, x)$, $s(x_r, x)$, α, β, γ represent the luminance, contrast and structure information and their relative importance, respectively. μ_{x_r}, μ_x are the means of x_r and x and σ_{x_r}, σ_x are the standard deviations of x_r and x. $\sigma_{x_r x}$ is the co-variance of x_r and x and $C_1 = 0.01^2$, $C_2 = 0.03^2$, $C_3 = C_2/2$ are the fixed hyper-parameters.

From Scale 1 to the highest Scale M, we compute the contrast comparison $c(x_r, x)$ and the structure comparison $s(x_r, x)$ at each scale j. The luminance comparison denoted as $l_M(x_r, x)$ is computed only at Scale M. M is set to 5 in our experiments.

Thus, we defined MS-SSIM [162] as

$$\begin{aligned} L_{MS-SSIM}(x_r, x) =& [l_M(x_r, x)]^{\alpha_M} \\ & * \prod_{i=1}^{M} [c(x_r, x)]^{\beta_i} [s(x_r, x)]^{\gamma_i}. \end{aligned} \tag{5.12}$$

Full SP-StarGAN loss function.

Finally, the proposed full generator loss function to optimize can be defined as

$$\begin{aligned} L_G = & L_{adv} + \lambda_{cls} L^f_{cls} + \lambda_{cls} L_{rec} \\ & + \lambda_{id} L_{id} - \lambda_s L_{MS-SSIM} \end{aligned} \tag{5.13}$$

where λ_{cls}, λ_{rec}, λ_{id}, λ_s are the relative importance of domain classification, reconstruction, identity mapping and MS-SSIM losses, respectively. Note that we have the same discriminator loss function as in Equation 5.5.

Network Architecture.

We employ the same network architecture from [42]. The generator is consist of two convolutional layers with stride size 2, 6 residual blocks [110] and two transposed

convolutional layers with stride size 2. The instance normalization [163] was used only for the generator. For the discriminator, the PatchGANs [164] was used to classify the local image patches are real or fake.

5.1.3 Deep Person ReID Network

Base Deep ReID Model.

We use the ID-Discriminative Embedding (IDE) [125] to train ReID model. In this network, we use ResNet-50 [110] convolutional layers followed by global max pooling layer. We then add two fully connected layers as stated in [43]. The first layer has 1024 dimensions followed by batch normalization [165], ReLU and Dropout [157]. For the ID-Discriminative Embedding, we have the second layer that has P (the number of class dimensions) in order to use cross-entropy loss.

Loss Function.

We use the cross-entropy loss for the real images. The cross-entropy loss is defined as

$$L_R = -\sum_{c=1}^{C} log(p(c))q(c) \tag{5.14}$$

where $p(c)$ is the estimated probability of the input with the ground truth label c and C is the number of classes. $q(c)$ the ground truth distribution and is defined as

$$q(c) = \begin{cases} 1, & c = y \\ 0, & c \neq y \end{cases} \tag{5.15}$$

For the generated images, we utilize the label smoothing regularization (LSR) as suggested in [43] to reduce the negative effect of some of the noisy generated images. The label smoothing regularization (LSR) loss can be defined as

$$L_{LSR} = -(1-\epsilon)\, log\, p(y) - \frac{\epsilon}{C}\sum_{c=1}^{C} log\, p(c) \tag{5.16}$$

Even though we have the identity label for the generated images, some images have transfer noise due to the occlusions or the noise in the input image. To alleviate this

problem, LSR assigns the small weights to the other classes and give less confidence in the identity label. The final loss of the Deep Person ReID Network is formulated as

$$L_T = \frac{1}{M}\sum_{i=1}^{M} L_R^i + \frac{1}{N}\sum_{j=1}^{N} L_{LSR}^j \tag{5.17}$$

where L_R is defined in Equation 5.14 and L_{LSR} is defined in Equation 5.16. M, N indicate the number of real images and generated images used in the training process.

Re-Rank.

We employ the re-ranking method [44] as post processing on our initial ranking results from base deep ReID model. Zhong *et al.* proposed to use the k-reciprocal encoding for ReID re-ranking. Re-rank computes features by encoding its k-reciprocal neighbors into a single vector. Then this vector is used to re-rank under the Jaccard distance. And the final distance is computed with the combination of the original distance and the Jaccard distance. We will refer this method as Re-Rank in the rest of the thesis.

5.2 Experiment Results

5.2.1 Datasets

Market-1501 [38] contains 32,668 images in total with 1,501 identities from 6 different camera views. From the video, person images were detected using a deformable part model [166]. This dataset is partitioned into 12,935 images (751 identities) for training and 19,732 images (750 identities) for the gallery. In ReID test, 3,3668 hand-captured images from 750 identities are pre-selected as queries to evaluate ReID performance. Single-query evaluation protocol is used.

DukeMTMC-reID [134] has 36,411 images in total with 1,404 identities from 8 different camera views. It is composed of 16,522 images (702 identities) for training samples and 17,661 images (702 identities) for the gallery. In ReID test, 2,228 images from 702 identities are pre-selected as queries for the evaluation.

Table 5.1.: ReID accuracy evaluation on different proposed components in SP-StarGAN loss on Market-1501

Method	λ_{id}	λ_s	mAP	Top-1 Rank
Baseline (IDE*)	-	-	65.87	85.66
StarGAN	0	0	66.1	86.5
StarGAN + Identity	1	0	67.2	86.7
	2	**0**	**68.2**	**87.9**
	5	0	67.4	87.5
StarGAN + MS-SSISM	0	1	67.4	87.4
	0	**2**	**67.6**	**87.4**
	0	5	66.2	85.9
StarGAN + Both	1	1	67	87.2
	1	2	67.6	87.5
	1	5	67.6	86.9
	2	**1**	**68.6**	**88.1**
	2	2	68.5	87.6
	2	5	67.6	87.6

5.2.2 Experiment Setup and Training Details

Similarity Preserving StarGAN. We first resize the images to 178x178 and then crop them randomly to 128x128. The horizontal random flip is used with a probability of 0.5 as the data augmentation. As described in [42], all models are trained using Adam [112] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The generator updates after five discriminator updates as in [159]. The initial learning rate is 0.0001 for the first 100,000 iterations and linearly decays to the learning to 0 over the next 100,000 iterations. The batch size is 16. We use the fixed hyper-parameter values for

Table 5.2.: ReID accuracy evaluation on different proposed components in SP-StarGAN loss on DukeMTMC-reID

Method	λ_{id}	λ_s	mAP	Top-1 Rank
Baseline (IDE*)	-	-	51.83	72.31
StarGAN	0	0	66.1	86.5
StarGAN + Identity	1	0	51.7	74.2
	2	0	52.4	74
	5	0	51.4	73.7
StarGAN + MS-SSISM	0	2	51.3	72.7
	0	2	49.6	71
	0	5	51.9	74
	1	1	51.9	73.9
StarGAN + Both	1	2	51.8	74.4
	1	5	51	72.9
	2	1	51.7	72.8
	2	2	51.7	73.6
	2	5	51.6	73.4

$\lambda gp = 10$, $\lambda cls = 1$, $\lambda rec = 10$ in Equation 5.13. We describe the analysis to select the best value for λ_{id}, λ_s in Section 5.2.3.

Finally, for inference, we generate all combinations of different cameras per image with the image size 128x128. For example, if the real image is taken from camera 1 and we have K different cameras in the dataset, then generate the translated images with target camera domain label from 2 to K.

Deep Person ReID Network. We follow the training general strategy in [43] to train the base deep reID model except for the learning rate policy. All images are resized to 256x128. Two base data augmentation method were used for the training : random cropping and random horizontal flipping. A model is trained with SGD

Table 5.3.: ReID accuracy evaluation on different proposed pre/post processing methods on Market-1501

Component	Base Augmentation		Base + RE		Base + Re-Rank		Base + RE + Re-Rank	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
StarGAN	66.1	86.5	69.8	88.4	82.1	88.9	85.6	90.9
StarGAN + Identity	68.2	87.9	71	88.9	83.3	89.8	86.7	91.3
StarGAN + MS-SSIM	67.6	87.4	71	88.7	82.8	89.3	85.6	91.2
StarGAN + Both	68.6	88.1	70.9	88.5	83	89.5	86.3	91.1

Table 5.4.: ReID accuracy evaluation on different proposed pre/post processing methods on DukeMTMC-reID

Component	Base Augmentation		Base + RE		Base + Re-Rank		Base + RE + Re-Rank	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
StarGAN	51.4	73.4	54.3	74.6	68.9	78.4	73.4	81.2
StarGAN + Identity	52.4	74	56.1	77	71	79.4	74.3	81.1
StarGAN + MS-SSIM	51.3	72.7	54.8	75	67.6	76.8	73.1	80.3
StarGAN + Both	51.7	72.8	55.2	75.9	70	79.4	65	82.1

solver. The initial learning rate is set to 0.01 for the ResNet-50 convolutional layers and 0.1 for the two additional fully connected layer since we use ImageNet [101] pre-trained ResNet-50 layers as the initialization. In our experiments, the initial learning is divided by 10 after first 30 epochs out of 60 epochs in total. The batch size is set to 128 and the dropout probability is set to 0.5.

In the ReID test, we extract the feature from the pooling layer and use Euclidean distance to compute the similarity between the gallery and query images. We use the generated images as the extra training samples and follow the strategy of [43] in the selection of the generated images. We randomly select M real images and N generated images in a training mini-batch. We set the $M : N$ ratio to $3 : 1$ for all experiments. We evaluate the ReID performance in terms of mean Average Precision (mAP) and Top-1 Rank matching accuracy.

5.2.3 Component Evaluation

In this section, we investigate the significance of the components in GAN part and ReID Network Part of the proposed method.

Similarity Preserving StarGAN.

We first investigate the effect of the additional loss terms in SP-StarGAN on ReID accuracy metrics. We evaluate for different hyper-parameter settings such as StarGAN + identity, StarGAN + MS-SSIM and StarGAN + Both when λ_{id} and λ_s are varying from $1 - 5$. Note that this evaluation was done without any additional augmentation or post-processing in Deep Re-ID network. In [43], they defined IDE* with the improved learning rate policy while keeping the same network architecture from IDE [125]. IDE* is used as the baseline to evaluate the proposed components. As shown in Table 5.2, the usage of original StarGAN [42] improved around 1% from the baseline in both mAP and Top-1 Rank accuracy. When we included the additional loss terms into the generator loss function, we obtain around 2% improvement in ReID accuracy depending on the hyper-parameters λ_{id} and λ_s. This improvement

is coming from the generating better quality images which results that having less noise in generated samples. We also observe that we do not have the continuing improvement as we increase the contributions of the additional loss terms.

Deep Person ReID Network.

We evaluate the different components in Deep ReID network including Random Erasing (RE) and Re-Rank. For this evaluation, we fix the hyper-parameters for the GAN part as $\lambda_{id} = 2$ and $\lambda_s = 1$. For any type of proposed GANs, we observe the significant improvement in ReID accuracy by employing both RE and Re-Rank. This result demonstrates that using Random Erasing as extra data augmentation along with the Re-Rank as the post processing has significant positive effect on ReID accuracy. Thus, our final proposed method version in the following section will be including both RE and Re-Rank as well as StarGAN + Both method with the parameters $\lambda_{id} = 2$ and $\lambda_s = 1$.

5.2.4 Complexity Analysis

Table 5.5 shows the comparison of the complexity of the model between CamStyle [43] and proposed method. Note that this experiment was done using a NVIDIA Titan Xp GPU. CamStyle has around 792 M parameters to train while our proposed method has only 52.23 M parameters to train as shown in Table 5.5a. For training and inference processing time as shown in Table 5.5b, CamStyle takes around 150 more hours in training than the proposed method for DukeMTMC-reID [37] dataset. Camstyle can only learn the mapping between two different camera domains at one time due to the limitation of CycleGAN. This results the dramatic increase in the complexity since we need to train multiple models. On the other hand, proposed method can model the mapping between multiple camera domains with the single model while showing the competitive ReID accuracy.

Table 5.5.: A complexity comparison on CamStyle [43] and Our Proposed Method

(a) Number of Parameters on DukeMTMC-reID [37]

Sub-Network	Number of Parameters [M]	
	CamStyle	Ours
Generator	637.17 M	8.44 M
Discriminator	154.84 M	44.79 M
Total	792.01 M	53.23 M

(b) Processing Time on DukeMTMC-reID [37]

Mode	Processing Time [hours]	
	CamStyle	Ours
Training	304.17	12.84
Inference	2.16	0.12

5.2.5 Comparisons

Visual Evaluation Comparison

We compare the sample generated images from Camstlyle and our proposed method. Both Camstlye and our proposed method can generate competitive quality of person images. However, as shown in Figure 5.2, in this particular sample, proposed method can generate better quality images especially in person's leg compare to Camstyle [43] and the original StarGAN [42]. This particular sample has a lot of noise in the input image and it demonstrates that proposed method can create better quality image even with the noisy input.

ReID Evaluation Comparison

For the full version of proposed method, we use the StarGAN + Both where $\lambda_{id} = 2$ and $\lambda_s = 1$ with RE and Re-Rank. We compare our proposed method with the state-of-the-art methods on Market-1501 and DukeMTMC-reID in Table 5.6 and 5.7. In

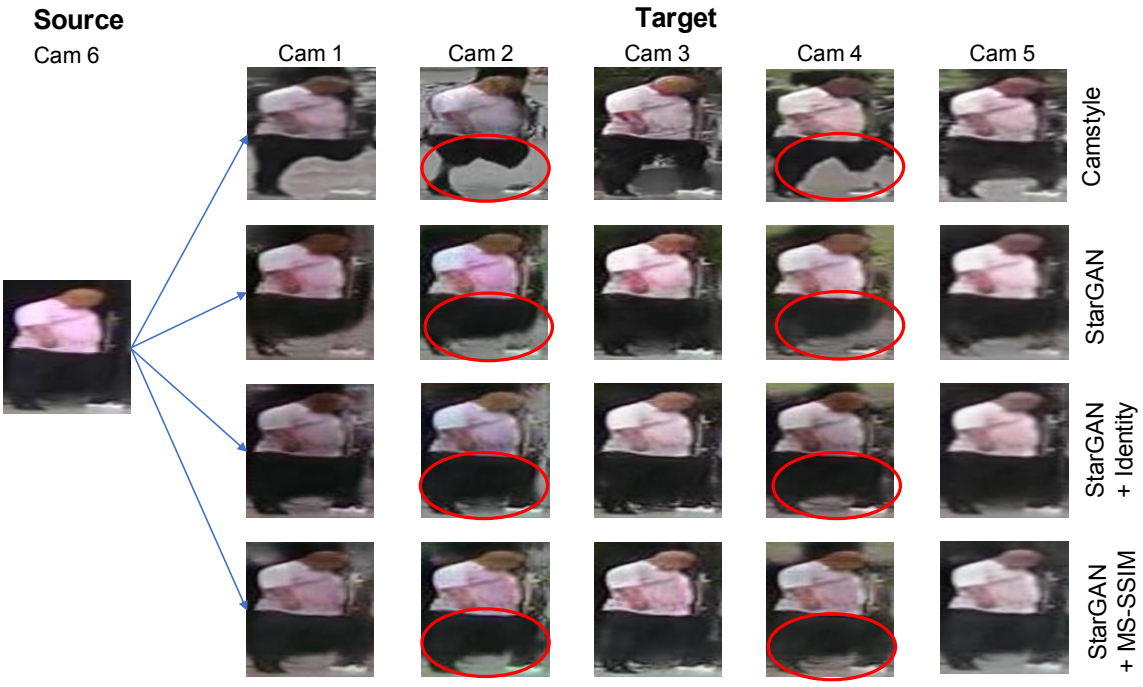


Fig. 5.2.: Sample Generated Image Comparison

both datasets, our proposed method outperforms all the other methods in terms of both mean Average Precision(mAP) and Top-1 Rank accuracy. mAP can be defined as

$$mAP = \frac{\sum_{q=1}^{Q}(\sum_{k=1}^{N} P_q(k)\Delta r_q(k))}{Q} \tag{5.18}$$

where $P_q(k)$ is the precision at a cutoff of k images given query q and $\Delta r_q(k)$ is the change in recall which happened between $k-1$ and k cutoff given query q.

We achieve significant improvement in especially mAP (15-72%) by employing Re-Rank with SP-StarGAN. We also achieve the highest accuracy in terms of Top-1 Rank accuracy in both datasets.

Fig. 5.3.: Sample Generated Image Comparison

Table 5.6.: A ReID accuracy comparison on Market-1501

Methods	mAP	Top-1 Rank
LOMO + XQDA [41]	14.09	34.4
IDE [125]	46	72.54
Re-rank [44]	63.63	77.11
SVDNet [78]	62.1	82.3
TriNet [167]	69.14	84.92
DJL [168]	65.5	85.1
DCGAN [37]	66.07	83.97
IDE* [43]	65.87	85.66
IDE* + CamStyle [43]	68.72	88.12
IDE* + CamStyle + RE [134]	71.55	89.49
Ours (full version)	86.3	91.1

Table 5.7.: A ReID accuracy comparison on DukeMTMC-reID

Methods	mAP	Top-1 Rank
BOW + KISSME [68]	12.17	25.13
LOMO + XQDA [41]	17.04	30.75
IDE [125]	44.99	65.22
SVDNet [78]	56.8	76.7
TriNet [167]	72.44	53.5
DCGAN [37]	47.13	67.68
IDE * [43]	51.83	72.31
IDE * + CamStyle [43]	53.48	75.27
IDE* + CamStyle + RE [134]	57.61	78.32
Ours (full version)	65	82.1

6. CONCLUSIONS

6.1 Summary

In this thesis, we describe improved video heart rate estimation method and two person re-identification methods. The main contributions of this thesis are listed as follows:

- Improving Video Heart Rate Estimation

    1. We reviewed the literature of video-based HR estimation methods. There were Independent Component Analysis (ICA) Approach, Motion Detection/Amplification Approach, Chrominance Based Approach and Other Approaches. We also stated the problem of current level of work. VHR method still have rooms for improvement by solving motion artifacts or varying skin tone problems.

    2. We propose a new VHR method using temporal differencing filter and small variation amplification. Since the definition of HR is the small color variation underneath the skin in our application, we employ the recursive temporal differencing filter to get temporal variations. We then perform small variation amplification using Equation 3.4 and 3.3 that amplifies only small variation and reduce large variation. To get more stable estimations, we also propose reduction of frequency range for bandpass filter using a cutoff frequency search.

    3. For the future direction, we propose an alternative method of skin detection, spatial pruning, for noise signal removal. By using histogram and K-means clustering within the tracked face, we expect to get more stable and accurate information for the HR estimation. We provide the justifi-

cation of illumination incident angle effect on our HR signal. This effect is shown as motion artifact since it is changing along the subject's motion. To alleviate motion artifact problem from illumination, we plan to implement the separation of illumination and reflectance from the image by using homomorphic filter.

- Person Re-Identification using a Two Stream Siamese CNN

    1. We propose a person re-identification method based on a two stream convolutional neural network where each stream is a Siamese network. This architecture can learn spatial and temporal information separately in a re-identification setting. By having two separate networks, each network can learn its own best feature representation.
    2. We propose a weighted two stream training objective function which combines the Siamese cost of the spatial and temporal streams with the objective to predict a person's identity. The weighted cost function controls the individual contribution of each stream.
    3. We evaluate our proposed method on two publicly available datasets. Our experimental results also demonstrate that by having two separate networks to represent the spatial and the temporal content, each network is able to learn the best feature representation and improves the ReID performance. Our proposed method outperforms or shows comparable results to the existing best perform methods on two public datasets.

- Person Re-Identification using a Similarity Preserving StarGAN (SP-StarGAN)

    1. We propose the scene-aware multiple domain image-to-image translation using Similarity Preserving StarGAN for Person Re-Identification. SP-StarGAN has the identity mapping loss and Multi-scale Structural Similarity loss in the generator loss function. The SP-StarGAN can learn the mapping among all different scene settings in ReID dataset and generate

the scene-aware translated images as the extra training samples in ReID with a single model.

2. For ReID, we propose to employ the Re-Ranking method [44] as post processing along with SP-StarGAN generated samples in order to improve ReID matching accuracy. We empirically demonstrate that Re-Ranking shows higher performance in ReID accuracy with better quality generated images.

3. We also provide the experimental results showing that having two additional loss terms helps address the quality problem in generated images as well as ReID performance. We also demonstrate that by using SP-StarGAN along with Random Erasing and Re-Rank improves the ReID performance.

6.2 Future Work

Our proposed methods can be extended in the following ways:

- Improving Video Heart Rate Estimation

  In the section 3.4 we have shown how the motion artifact is involved with the heart rate signal through the mathematical modeling. Since the nature of illumination in image capture is multiplicative model, the illumination produces modulation artifacts term in the signal as shown in Equation 3.20. The experiments on the synthetic data show that the modulation artifacts overwhelm our target signal (HR signal) and reduce the estimation accuracy.

  To overcome this problem, we plan to have an improved system for video based-HR estimation by implementing the future work ideas described in Section 3.3 and 3.4. The overall future block diagram is given in Figure 6.1 where homomorphic filter will be employed before Spatial Pruning.

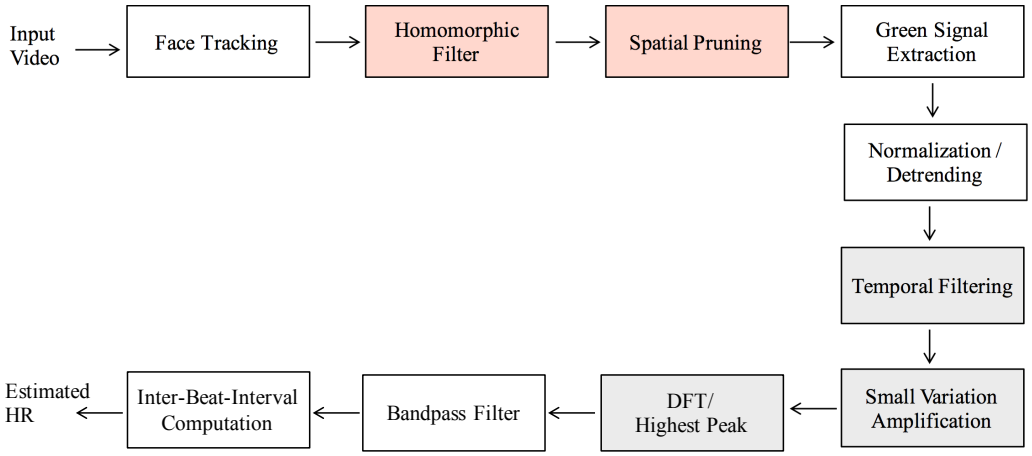


Fig. 6.1.: The overall block diagram of the future work.

As Figure 6.1 shows, we propose to use homomorphic filtering to remove the illumination effect from the given image. In addition, we plan to implement 2D homomorphic filtering on image domain for the purpose of separating illumination from image. Furthermore, we plan to analyze how this idea will actually affect on real video data and HR estimation.

- Person Re-Identification using a Two Stream Siamese CNN

  In our work, we only consider the spatial and temporal information of the person. However, semantic attributes are critical and robust features to identify a person. To address this, in the future, we plan to incorporate semantic attributes using a multi-stream approach to address the challenges associated with occlusions and cluttered background.

- Person Re-Identification using a Similarity Preserving StarGAN (SP-StarGAN)

  Even though most of the person re-identification (ReID) methods has competitive accuracy, it still works only well on same dataset as the trainig dataset. In the real-world setting, we expect the method to be able to identify a person even if the person is not coming from similar environmental setting to the training data. This is a challenging task and cross-dataset domain ReID is another pop-

ular research area. In the future, we want to extend the usage of SP-StarGAN to cross-dataset domain ReID problem.

6.3 Publications Resulting From Our Work

1. **D. Chung**, and E. J. Delp, "Camera-Aware Image-to-Image Translation Using Similarity Preserving StarGAN For Person Re-identification," *To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2019, Long Beach, CA.

2. **D. Chung**, K. Tahboub and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," *Proceedings of the International Conference on Computer Vision*, pp. 1983-1991, October 2017, Venice, Italy.

3. **D. Chung**, J. Choe, M. E. OHaire, A.J. Schwichtenberg, and E. J. Delp, "Improving video-based heart rate estimation," *Proceedings of the IS&T International Symposium on Electronic Imaging*, February 2016, San Francisco, CA.

4. J. Choe, **D. Chung**, A. J. Schwichtenberg, and E. J. Delp, "Improving video-based resting heart rate estimation: A comparison of two methods," *Proceedings of the IEEE 58th International Midwest Symposium on Circuits and Systems*, pp. 1-4, August 2015, Fort Collins, CO.

REFERENCES

REFERENCES

[1] J. Achten and A. Jeukendrup, "Heart rate monitoring," *Sports medicine*, vol. 33, no. 7, pp. 517–538, 2003. [Online]. Available: http://dx.doi.org/10.2165/00007256-200333070-00004

[2] T. Taylor, "cardiovascular system," https://www.innerbody.com/image/cardov.html.

[3] L. J. Vorvick, "U.s. national library of medicine, pulse," https://www.nlm.nih.gov/medlineplus/ency/article/003399.htm.

[4] R. A. Robergs and R. Landwehr, "The surprising history of the "hrmax= 220-age" equation," *Journal of Exercise Physiology*, vol. 5, no. 2, pp. 1–10, 2002.

[5] C. Brüser, S. Winter, and S. Leonhardt, "Robust inter-beat interval estimation in cardiac vibration signals," *Physiological measurement*, vol. 34, no. 2, p. 123, 2013. [Online]. Available: http://dx.doi.org/10.1088/0967-3334/34/2/123

[6] B.-U. Köhler, C. Hennig, and R. Orglmeister, "The principles of software qrs detection," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, no. 1, pp. 42–57, 2002. [Online]. Available: http://dx.doi.org/10.1109/51.993193

[7] W. Verkruysse, L. Svaasand, and J. Nelson, "Remote plethysmographic imaging using ambient light," *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[8] A. Hertzman and C. Spealman, "Observations on the finger volume pulse recorded photoelectrically," *Am. J. Physiol*, vol. 119, no. 334, p. 3, 1937.

[9] A. Challoner, "Photoelectric plethysmography for estimating cutaneous blood flow," *Non-invasive physiological measurements*, vol. 1, pp. 125–51, 1979.

[10] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, pp. R1–R39, March 2007. [Online]. Available: http://dx.doi.org/10.1088/0967-3334/28/3/R01

[11] S. Kim, D. Ryo, and C. Bae, "Adaptive noise cancellation using accelerometers for the ppg signal from forehead," *Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society*, pp. 2564–2567, Aug 2007, Lyon. [Online]. Available: http://dx.doi.org/10.1109/IEMBS.2007.4352852

[12] S. Seyedtabaii and L. Seyedtabaii, "Kalman filter based adaptive reduction of motion artifact from photoplethysmographic signal," *World Academy of Science, Engineering and Technology*, vol. 37, pp. 173–176, 2008.

[13] C. Lee and Y. Zhang, "Reduction of motion artifacts from photoplethysmographic recordings using a wavelet denoising approach," *Proceedings of the IEEE EMBS Asian-Pacific Conference on Biomedical Engineering, 2003*, pp. 194–195, October 2003. [Online]. Available: http://dx.doi.org/10.1109/APBME.2003.1302650

[14] M. Garbey, N. Sun, A. Merla, and I. Pavlidis, "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 8, pp. 1418–1426, Aug 2007. [Online]. Available: http://dx.doi.org/10.1109/TBME.2007.891930

[15] M. PDoi and S. Tominaga, "Spectral estimation of human skin color using the kubelka-munk theory," *Proceedings of the Electronic Imaging conference on Color Imaging VIII: Processing, Hardcopy, and Applications*, pp. 221–228, Januray 2003, Santa Clara,CA. [Online]. Available: http://dx.doi.org/10.1117/12.472026

[16] N. Haering, P. L. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," *Machine Vision and Applications*, vol. 19, no. 5, pp. 279–290, October 2008. [Online]. Available: https://doi.org/10.1007/s00138-008-0152-0

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

[18] M. Shah, O. Javed, and K. Shafique, "Automated visual surveillance in realistic scenarios," *IEEE MultiMedia*, vol. 14, no. 1, pp. 30–39, Jan 2007. [Online]. Available: https://doi.org/10.1109/MMUL.2007.3

[19] H. Dee and S. Velastin, "How close are we to solving the problem of automated visual surveillance?" *Machine Vision and Applications*, vol. 19, no. 329, 2008. [Online]. Available: https://doi.org/10.1007/s00138-007-0077-z

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. I–511–I–518, December 2001, Kauai, HI. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2001.990517

[21] R. Girshick, "Fast r-cnn," *Proceedings of the International Conference on Computer Vision*, pp. 1440–1448, Dec 2015. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.169

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, June 2016. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.91

[23] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *Proceedings of the IEEE International Conference on Image Processing*, pp. 3645–3649, Sep 2017. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2017.8296962

[24] Y. Wu, J. Lim, and M. Yang, "Online object tracking: A benchmark," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411–2418, June 2013. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2013.312

[25] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *Proceedings of the ACM International Conference on Multimedia*, pp. 357–360, 2007, Augsburg, Germany. [Online]. Available: http://doi.acm.org/10.1145/1291233.1291311

[26] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Proceedings of the Advances in Neural Information Processing Systems*, pp. 568–576, December 2014, Montreal, Canada. [Online]. Available: http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *Proceedings of the International Conference on Computer Vision*, pp. 4489–4497, 2015, Washington, DC. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.510

[28] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1933–1941, June 2016. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.213

[29] M. Butenuth, F. Burkert, F. Schmidt, S. Hinz, D. Hartmann, A. Kneidl, A. Borrmann, and B. Sirmacek, "Integrating pedestrian simulation, tracking and event detection for crowd analysis," *Proceedings of the International Conference on Computer Vision Workshops*, pp. 150–157, Nov 2011. [Online]. Available: https://doi.org/10.1109/ICCVW.2011.6130237

[30] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," *Machine Vision and Applications*, p. 345, 2008. [Online]. Available: https://doi.org/10.1007/s00138-008-0132-4

[31] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 15:1–15:58, july 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541882

[32] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, June 2018, Salt Lake City, UT. [Online]. Available: http://doi.acm.org/10.1109/CVPR.2018.00678

[33] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, April 2014. [Online]. Available: https://doi.org/10.1016/j.imavis.2014.02.001

[34] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. London: Springer, 2014. [Online]. Available: https://doi.org/10.1007/978-1-4471-6296-4

[35] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," *Proceedings of the Scandinavian Conference on Image Analysis*, pp. 91–102, May 2011, Ystad, Sweden. [Online]. Available: https://doi.org/10.1007/978-3-642-21227-7_9

[36] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," *Proceedings of the European Conference on Computer Vision*, pp. 688–703, September 2014, Zurich, Switzerland. [Online]. Available: https://doi.org/10.1007/978-3-319-10593-2_45

[37] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *Proceedings of the IEEE International Conference on Computer Vision*, July 2017, Honolulu, Hawaii. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.405

[38] L. Zheng, 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*, December 2015, Las Condes, Chile. [Online]. Available: https://doi.org/10.1109/ICCV.2015.133

[39] S. G. S, M. C. M, C.C.Loy, and T. Hospedales, *The Re-identification Challenge*. London: Springer, 2014. [Online]. Available: https://doi.org/10.1007/978-1-4471-6296-4_1

[40] R. Layne, T. M. Hospedales, and S. Gong, "Domain transfer for person re-identification," *Proceedings of the ACM/IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream*, pp. 25–32, 2013, Barcelona, Spain. [Online]. Available: http://doi.acm.org/10.1145/2510650.2510658

[41] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197–2206, June 2015, Boston, MA. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298832

[42] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, Salt Lake City, UT. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00916

[43] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, Salt Lake City, UT. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00541

[44] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," *Proceedings of the IEEE International Conference on Computer Vision*, July 2017, Honolulu, Hawaii. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.389

[45] Y. Sun, C. Papin, V. Azorin-Peris, R. Kalawsky, S. Greenwald, and S. Hu, "Use of ambient light in remote photoplethysmographic systems: comparison between a high-performance camera and a low-cost webcam," *Journal of biomedical optics*, vol. 17, no. 3, pp. 037 005 1–037 005 10, 2012. [Online]. Available: http://dx.doi.org/10.1117/1.JBO.17.3.037005

[46] M. Poh, D. McDuff, and R. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Express*, vol. 18, no. 10, pp. 10 762–10 774, May 2010. [Online]. Available: http://dx.doi.org/10.1364/OE.18.010762

[47] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000. [Online]. Available: https://doi.org/10.1016/S0893-6080(00)00026-5

[48] M. Poh, D. McDuff, and R. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, January 2011. [Online]. Available: http://dx.doi.org/10.1109/TBME.2010.2086456

[49] M. Tarvainen, P. Ranta-aho, and P. Karjalainen, "An advanced detrending method with application to hrv analysis," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 2, pp. 172–175, February 2002. [Online]. Available: http://dx.doi.org/10.1109/10.979357

[50] J. Vila, F. Palacios, J. Presedo, M. Fernandez-Delgado, P. Felix, and S. Barro, "Time-frequency analysis of heart-rate variability," *IEEE Magazine on Engineering in Medicine and Biology*, vol. 16, no. 5, pp. 119–126, September/October 1997. [Online]. Available: http://dx.doi.org/10.1109/51.620503

[51] H. Monkaresi, R. Calvo, and H. Yan, "A machine learning approach to improve contactless heart rate monitoring using a webcam," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, pp. 1153–1160, November 2013. [Online]. Available: http://dx.doi.org/10.1109/JBHI.2013.2291900

[52] D. McDuff, S. Gontarek, and R. Picard, "Improvements in remote cardiopulmonary measurement using a five band digital camera," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2593 – 2601, October 2014. [Online]. Available: http://dx.doi.org/10.1109/TBME.2014.2323695

[53] J. Choe, D. Chung, A. J. Schwichtenberg, and E. J. Delp, "Improving video-based resting heart rate estimation: A comparison of two methods," *Proceedings of the IEEE 58th International Midwest Symposium on Circuits and Systems*, pp. 1–4, August 2015, Fort Collins, CO. [Online]. Available: http://dx.doi.org/10.1109/MWSCAS.2015.7282155

[54] H. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 65:1–8, July 2010. [Online]. Available: http://dx.doi.org/10.1145/2185520.2185561

[55] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983. [Online]. Available: http://dx.doi.org/10.1109/TCOM.1983.1095851

[56] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3430–3437, June 2013, Portland, OR. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2013.440

[57] V. J. G. de Haan, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013. [Online]. Available: http://dx.doi.org/10.1109/TBME.2013.2266196

[58] G. de Haan and A. Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, no. 9, pp. 1913–1926, 2014. [Online]. Available: http://dx.doi.org/10.1088/0967-3334/35/9/1913

[59] W. Wang, S. Stuijk, and G. de Haan, "Exploiting spatial redundancy of image sensor for motion robust rppg," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 415–425, Feb 2015. [Online]. Available: http://dx.doi.org/10.1109/TBME.2014.2356291

[60] H. E. Tasli, A. Gudi, and M. den Uyl, "Remote ppg based vital sign measurement using adaptive facial regions," *Proceedings of the IEEE International Conference on Image Processing*, pp. 1410–1414, Oct 2014, France, Paris. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2014.7025282

[61] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 300–305, April 1998, nara. [Online]. Available: http://dx.doi.org/10.1109/AFGR.1998.670965

[62] L. Feng, L.-M. Po, X. Xu, Y. Li, C.-H. Cheung, K.-W. Cheung, and F. Yuan, "Dynamic roi based on k-means for remote photoplethysmography," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1310–1314, April 2015, South Brisbane, QLD. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2015.7178182

[63] L. Feng, L. Po, X. Xu, Y. Li, and R. Ma, "Motion-resistant remote imaging photoplethysmography based on the optical properties of skin," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 879–891, 2015. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2014.2364415

[64] Y. Yan, X. Ma, L. Yao, and J. Ouyang, "Non-contact measurement of heart rate using facial video illuminated under natural light and signal weighted analysis," *Bio-Medical Materials and Engineering*, vol. 26, no. s1, pp. 903–909, 2015. [Online]. Available: http://dx.doi.org/10.3233/BME-151383

[65] M. Kumar, A. Veeraraghavan, and A. Sabharval, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–1588, 2015. [Online]. Available: http://dx.doi.org/10.1364/BOE.6.001565

[66] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *Proceedings of the 10th European Conference on Computer Vision*, pp. 262–275, October 2008, Marseille, France. [Online]. Available: https://doi.org/10.1007/978-3-540-88682-2_21

[67] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1363–1372, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.152

[68] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288–2295, June 2012, Providence, RI. [Online]. Available: https://doi.org/10.1109/CVPR.2012.6247939

[69] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," *Proceedings of the 13th European Conference on Computer Vision*, pp. 1–16, October 2014, Zurich, Switzerland. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10584-0_1

[70] R. Layne, T. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," *Proceedings of the British Machine Vision Conference*, vol. 2, no. 3, p. 8, September 2012, Guildford, United Kingdom. [Online]. Available: http://dx.doi.org/10.5244/C.26.24

[71] R. Layne, T. M. T. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," *Proceedings of the European Conference on Computer Vision*, pp. 402–412, October 2012, Berlin, Heidelberg. [Online]. Available: https://doi.org/10.1007/978-3-642-33863-2_40

[72] S. Khamis, C. Kuo, V. Singh, V. Shet, and L. Davis, "Joint learning for attribute-consistent person re-identification," *Proceedings of the European Conference on Computer Vision Workshops*, pp. 134–146, October 2014, Zurich, Switzerland. [Online]. Available: https://doi.org/10.1007/978-3-319-16199-0_10

[73] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2666–2672, June 2012, Providence, RI. [Online]. Available: https://doi.org/10.1109/CVPR.2012.6247987

[74] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," *Proceedings of the 10th European Conference on Computer Vision*, pp. 780–793, October 2012, Florence, Italy. [Online]. Available: https://doi.org/10.1007/978-3-642-33783-3_56

[75] J. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," *Proceedings of the 24th International Conference on Machine Learning*, pp. 209–216, June 2007, Corvallis, OR. [Online]. Available: http://dx.doi.org/10.1145/1273496.1273523

[76] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," *Proceedings of the European Conference on Computer Vision*, October 2016, Amsterdam, Netherlands. [Online]. Available: https://doi.org/10.1007/978-3-319-46466-4_52

[77] D. Chung, K. Tahboub, and E. Delp, "A two stream siamese convolutional neural network for person re-identification," *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017, Venice, Italy. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.218

[78] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," *Proceedings of the IEEE International Conference on Computer Vision*, Oct

2017, Venice, Italy. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.410

[79] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, Salt Lake City, UT. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00242

[80] C. Schmid, "Constructing models for content-based image retrieval," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, Dec 2001. [Online]. Available: https://doi.org/10.1109/CVPR.2001.990922

[81] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biological Cybernetics*, vol. 61, no. 2, pp. 103–113, Jun 1989. [Online]. Available: https://doi.org/10.1007/BF00204594

[82] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and Z. Stan, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1301–1306, June 2010, San Francisco, CA. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2010.5539817

[83] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2002.1017623

[84] R. Satta, "Appearance descriptors for person re-identification: a comprehensive review," *ArXiv preprints*, 2013.

[85] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, Dec 2013. [Online]. Available: https://doi.org/10.1007/s11263-013-0636-x

[86] W. Li, R. Zhao, and X. W. Xiaogang, "Human re-identification with transferred metric learning," *Proceedings of the Asian Conference on Computer Vision*, pp. 31–44, 2013. [Online]. Available: http://doi.acm.org/110.1007/978-3-642-37331-2_3

[87] Y. Li, Z. Wu, and R. J. Radke, "Multi-shot re-identification with random-projection-based random forests," *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 373–380, January 2015, Waikoloa, HI. [Online]. Available: http://dx.doi.org/10.1109/WACV.2015.56

[88] W. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, March 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.138

[89] S. Karanam, Y. Li, and R. Radke, "Sparse re-id: Block sparsity for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 33–40, June 2015, Boston, MA. [Online]. Available: http://dx.doi.org/10.1109/CVPRW.2015.7301392

[90] K. Tahboub, B. Delgado, and E. J. Delp, "Person re-identification using a patch-based appearance model," *Proceedings of the IEEE Conference on Image Processing*, pp. 764–768, September 2016, Phoenix, AZ. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2016.7532460

[91] Y. Li, Z. Wu, S. Karanam, and R. Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis," *Proceedings of the British Machine Vision Conference*, September 2015, Swansea, United Kingdom. [Online]. Available: http://dx.doi.org/10.5244/C.29.73

[92] Z. Wu, Y. Li, and R. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1095–1108, September 2014. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2014.2360373

[93] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," MIT Press, 2016.

[94] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[95] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, p. 22222232, October 2017, Lake Tahoe, NV. [Online]. Available: https://doi.org/10.1109/TNNLS.2016.2582924

[96] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *ArXiv preprints*, 2015.

[97] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, p. 1842, July 2017. [Online]. Available: https://doi.org/10.1109/MSP.2017.2693418

[98] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proceedings of the Neural Information Processing Systems*, p. 10971105, December 2017, Lake Tahoe, NV. [Online]. Available: http://doi.acm.org/10.1145/3065386

[99] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998. [Online]. Available: https://doi.org/10.1109/5.726791

[100] R. Raina, A. Madhavan, and A. Y. Ng, "Large-scale deep unsupervised learning using graphics processors," *Proceedings of the Annual International Conference on Machine Learning*, pp. 873–880, 2009, Montreal, Quebec, Canada. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553486

[101] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848

[102] Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, M. andSean, H. andZhiheng, A. Karpathy, Khosla, Aditya, Bernstein, Michael, Berg, A. C., and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec 2015. [Online]. Available: https://doi.org/10.1007/s11263-015-0816-y

[103] M. Everingham, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, p. 303338, 2010. [Online]. Available: https://doi.org/10.1007/s11263-009-0275-4

[104] M. Everingham, S. M. A. Eslami, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, p. 98136, 2015. [Online]. Available: https://doi.org/10.1007/s11263-014-0733-5

[105] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, Columbus, OH.

[106] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Proceedings of the European Conference on Computer Vision*, pp. 818–833, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-10590-1_53

[107] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *Proceedings of the Deep Learning Workshop in International Conference on Machine Learning*, July 2015, Lille grand, Paris.

[108] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv preprints*, 2015.

[109] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298594

[110] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.90

[111] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec 1989. [Online]. Available: https://doi.org/10.1162/neco.1989.1.4.541

[112] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv preprints*, December 2014.

[113] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," *Proceedings of the BigLearn workshop at the Neural Information Processing Systems*, pp. 1–6, Dec 2011, Granada, Spain. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.231.4195

[114] M. Abadi, P. Barham, Z. C. J. Chen, A. Davis, M. D. J. Dean, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, , and X. Zheng, "Tensorflow: A system for large-scale machine learning," *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, p. 265283, November 2016, Savannah, GA. [Online]. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

[115] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," *Proceedings of the Autodiff Workshop at the Advances in Neural Information Processing Systems*, pp. 1–4, Dec 2017, Long Beach, CA. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ

[116] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *ArXiv preprints*, 2014.

[117] W. Li, 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, June 2014, Columbus, OH. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.27

[118] E. Ahmed, 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, June 2015, Boston, MA. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7299016

[119] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," *Proceedings of the International Conference on Pattern Recognition*, pp. 34–39, August 2014, Stockholm,Sweden. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2014.16

[120] 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," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335–1344, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.149

[121] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1249–1258, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.140

[122] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1288–1296, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.144

[123] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," *Proceedings of the European Conference on Computer Vision*, Oct 2016, Amsterdam, Netherlands. [Online]. Available: https://doi.org/10.1007/978-3-319-46478-7_9

[124] N. McLaughlin, J. Martinez, and P. Miller, "Recurrent convolutional network for video-based person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1325–1334, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.148

[125] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *ArXiv preprints*, 2016.

[126] J. Si, H. Zhang, C. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5363–5372, June 2018. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00562

[127] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Transactions on Multimedia*, vol. PP, 10 2018. [Online]. Available: https://doi.org/10.1109/TMM.2018.2877886

[128] E. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 1269–1277, 2014, montreal, Canada. [Online]. Available: http://dl.acm.org/citation.cfm?id=2968826.2968968

[129] S. Bai and X. Bai, "Sparse contextual activation for efficient visual re-ranking," *IEEE Transactions on Image Processing*, 2016. [Online]. Available: http://dx.doi.org/10.1109/TIP.2016.2514498

[130] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2672–2680, 2014, montreal, Canada. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets

[131] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Proceedings of the International Conference on Learning Representations*, vol. abs/1511.06434, 2016, Vancouver, Canada.

[132] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE International Conference on Computer Vision*, July 2017, Honolulu, Hawaii. [Online]. Available: http://dx.doi.org/10.1109/10.1109/CVPR.2017.632

[133] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017, Venice, Italy. [Online]. Available: http://dx.doi.org/10.1109/10.1109/ICCV.2017.244

[134] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *ArXiv preprints*, 2017.

[135] C. Zhang, L. Wu, and Y. Wangs, "Crossing generative adversarial networks for cross-view person re-identification," *Neurocomputing*, vol. 340, 01 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.01.093

[136] J. Lv and X. Wang, "Cross-dataset person re-identification using similarity preserved generative adversarial networks," *Knowledge Science, Engineering and Management*, pp. 171–183, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-99247-1_15

[137] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, Salt Lake City, UT. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00110

[138] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, Salt Lake City, UT.

[139] Z. Yin, W.-S. Zheng, A. Wu, H.-X. Yu, H. Wan, X. Guo, F. Huang, and J. Lai, "Adversarial attribute-image person re-identification," *International Joint Conferences on Artificial Intelligence*, 2018.

[140] B. Delgado, K. Tahboub, and E. J. Delp, "Superpixels shape analysis for carried object detection," *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, pp. 1–6, March 2016, Lake Placid, NY. [Online]. Available: http://dx.doi.org/10.1109/WACVW.2016.7470116

[141] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 869–878, May 2015. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2014.23525527

[142] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *ArXiv preprints*, 2017.

[143] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," *Proceedings of the European Conference on Computer Vision*, pp. 661–675, May 2002, Copenhagen, Denmark. [Online]. Available: http://dx.doi.org/10.1007/3-540-47977-5

[144] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2003.1195991

[145] D. Chai, S. Phung, and A. Bouzerdoum, "A bayesian skin/non-skin color classifier using non-parametric density estimation," *Proceedings of the International Symposium on Circuits and Systems*, vol. 2, pp. II–464–II–467, May 2003. [Online]. Available: http://dx.doi.org/10.1109/ISCAS.2003.1206010

[146] J. Casati, D. Moraes, and E. Rodrigues, "Sfa: A human skin image database based on feret and ar facial images," *Proceedings of the IX workshop de Visao Computational*, June 2013, rio de Janeiro, Brazil.

[147] J. McDonald, *Handbook of Biological Statistics : Paired T-Test.* Baltimore, Maryland: Sparky House Publishing, 2014, vol. 3.

[148] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, First Edition.* Addison-Wesley Longman Publishing Co., Inc., 2005, boston, MA, USA.

[149] H. Barrow and J. Tannenbaum, "Recovering intrinsic scene characteristics from images," *Computer Vision Systems, A. Hanson and E. Riseman*, no. 157, pp. 3–26, 1978.

[150] M. Doi and S. Tominaga, "Spectral estimation of human skin color using the kubelka-munk theory," *Proceedings of the SPIE Conference on Color Imaging VIII: Processing, Hardcopy, and Applications*, vol. 5008, pp. 221–228, January 2003, Santa Clara, CA. [Online]. Available: http://dx.doi.org/10.1117/12.472026

[151] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham Jr, "Nonlinear filtering of multiplied and convolved signals," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 3, pp. 437–466, 1968. [Online]. Available: http://dx.doi.org/10.1109/PROC.1968.6570

[152] D. Toth, T. Aach, and V. Metzler, "Illumination-invariant change detection," *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 3–7, April 2000, austin,TX. [Online]. Available: http://dx.doi.org/10.1109/IAI.2000.839561

[153] J. You, A. Wu, X. Li, and W. Zheng, "Top-push video-based person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1345–1353, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.150

[154] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 2, pp. 162–177, 2005. [Online]. Available: https://doi.org/10.1109/TPAMI.2005.39

[155] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 25–44, January 1994. [Online]. Available: https://doi.org/10.1142/S0218001493000339

[156] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 81, pp. 674–679, 1981, Vancouver, Canada.

[157] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, June 2014.

[158] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1988–1996, December 2014, Montreal, Canada.

[159] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *Proceedings of the International Conference on Machine Learning*, vol. 70, pp. 214–223, Aug 2017, Sydney, Australia. [Online]. Available: http://proceedings.mlr.press/v70/arjovsky17a.html

[160] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 5769–5779, 2017, long Beach, CA. [Online]. Available: http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans

[161] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *Proceedings of the International Conference on Learning Representations*, April 2017, toulon, France.

[162] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, pp. 1398–1402, Nov 2003. [Online]. Available: https://doi.org/10.1109/ACSSC.2003.1292216

[163] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *ArXiv preprints*, 2016.

[164] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," *Proceedings of the European Conference on Computer Vision*, Oct 2016, Amsterdam, Netherlands. [Online]. Available: https://doi.org/10.1007/978-3-319-46487-9_43

[165] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proceedings of the International Conference on Machine Learning*, pp. 448–456, Jul 2015, lille, France. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045167

[166] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sep 2010. [Online]. Available: https://doi.org/10.1109/TPAMI.2009.167

[167] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *ArXiv preprints*, 2017.

[168] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2194–2200, 2017, melbourne, Australia. [Online]. Available: http://dx.doi.org/10.24963/ijcai.2017/305

VITA

VITA

Dahjung Chung was born in South Korea. She received the B.S. in Electronic and Information Engineering from Ewha Womans University, Seoul, korea. She received her Master of Science in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea. Ms. Chung joined the Ph.D. program at the School of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana in August 2013. She worked at the Video and Image Processing Laboratory (VIPER) under the supervision of Professor Edward J. Delp. She was an Video Imaging Research Intern of Dolby Laboratory, Sunnyvale, CA in the summer of 2017 and Video Analytic Technology Software Intern of NVIDIA, Santa Clara, CA in the summer of 2018. Her research interests are computer vision, machine learning and deep learning. She is a student member of the IEEE and the IEEE Signal Processing Society.