THERMAL TO VISUAL: A CROSS-DOMAIN FACE RECOGNITION SYSTEM

A thesis submitted during 2023 to the University of Hyderabad in partial fulfillment of the award of a Ph.D. degree in Computer Science

by

YASWANTH GAVINI

SCHOOL OF COMPUTER & INFORMATION SCIENCES

UNIVERSITY OF HYDERABAD

(P.O.) CENTRAL UNIVERSITY

HYDERABAD - 500 046, INDIA

June 30, 2023

CERTIFICATE

This is to certify that the thesis entitled **"Thermal to Visual: A Cross-Domain Face Recognition System"** submitted by **Yaswanth Gavini** bearing **Reg. No. 15MCPC16** in partial fulfilment of the requirements for the award of **Doctor of Philosophy** in **Computer Science** is a bonafide work carried out by him under our supervision and guidance.

This thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other university or institution for the award of any degree or diploma. The student has the following publications before submission of the thesis for adjudication and has produced evidence for the same.

1. Gavini, Yaswanth, Arun Agarwal, and B. M. Mehtre. "Cross-Domain Face Recognition Using Dictionary Learning." *International Conference on Multi-disciplinary Trends in Artificial Intelligence*. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-33709-4_15
2. Gavini, Yaswanth, B. M. Mehtre, and Arun Agarwal. "Thermal to Visual Face Recognition using Transfer Learning." *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*. IEEE, 2019. https://doi.org/10.1109/ISBA.2019.8778474
3. Gavini, Yaswanth, Arun Agarwal, B. M. Mehtre. "Thermal to Visual Person Re-Identification Using Collaborative Metric Learning Based on Maximum Margin Matrix Factorization." *Pattern Recognition*. Volume 134, 2023, 109069, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2022.109069

Further, the student has passed the following courses towards the fulfilment of the coursework requirement for Ph.D.

Course Code	Name	Credits	Pass/Fail
CS 801	Data Structure and Algorithms	4	Pass
CS 802	Operating Systems and Programming	4	Pass
AI 820	Digital Image Processing	4	Pass
AI 851	Trends in Soft Computing	4	Pass

Prof. Arun Agarwal
Supervisor

Prof. B. M. Mehtre
Supervisor

30/6/23

Prof. Atul Negi
Dean, School of Computer and Information Sciences

DECLARATION

I, **Yaswanth Gavini**, hereby declare that this thesis entitled **"Thermal to Visual: A Cross-Domain Face Recognition System"** submitted by me under the guidance and supervision of **Prof. Arun Agarwal** and **Prof. B. M. Mehtre** is a bonafide research work. I also declare that it has not been submitted previously in part or in full to this university or any other university or institution for the award of any degree or diploma.

Date: 30-06-2023

Name: **Yaswanth Gavini**

Signature of the Student: G. Yaswanth

Reg. No. **15MCPC16**

This dissertation is dedicated to my family members and Teachers.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to the individuals who supported me in this academic journey. First and foremost, I express my sincere gratitude to my supervisors, **Prof. Arun Agarwal** and **Prof. B. M. Mehtre**, for their guidance and encouragement throughout this academic journey. Their valuable feedback and support have helped me grow academically and personally.

I would also like to thank my Doctoral Review Committee members, **Prof. C. R. Rao** and **Prof. Rajeev Wankar**, for their valuable insights and suggestions. Their thoughtful feedback and constructive criticism have been instrumental in shaping my work.

I express my gratitude to the Ministry of Electronics and Information Technology (MeitY), Government of India, for their funding of this research through the Visvesvaraya PhD Scheme, which is administered by Digital India Corporation (formerly Media Lab Asia).

Furthermore, I would like to thank my lab mates Rajesh, Rakesh, Renjith, Salman, Thirupathi, Akshay, N. D. Patel and Abhay for their support and valuable discussions. Their insights and perspectives have been invaluable in helping me to refine my research.

I would like to express my gratitude to my family members Satyanarayana, Udaya Sri, Jayanth, Swetha, Aditya, Karthikeya, Vaasudev Sai, Subba Rao, Kathyayani Devi and Vivek for their continuous support and encouragement.

Finally, I would like to thank my wife, Sowmini Devi, who has been a source of unwavering support and encouragement throughout my journey. Her faith in me has kept me grounded and motivated even during challenging times. I am incredibly fortunate to have such an exceptional partner in my life who always inspires me to enhance my skills.

Yaswanth Gavini

ABSTRACT

In recent years, automatic surveillance systems have evolved significantly. Face recognition is an essential task in many automatic surveillance systems. Face recognition in visual light images is a well-explored problem in computer vision. But most of the night-time security and military surveillance goes beyond visual domains such as near-infrared, thermal and Depth images. Among those, thermal images are captured from the natural thermal emission of the body. Because of this advantage, thermal images are more suitable for night-time surveillance applications. Many visual light face recognition algorithms are available to address the discrepancy due to illumination and pose problems. These traditional face recognition algorithms fail to handle the discrepancy due to domain differences.

Cross-domain face recognition algorithms have been introduced to address the discrepancy due to domain differences. Among the cross-domain face recognition problems, thermal to visual face recognition is more challenging because of the differences in spectral characteristics of the visual and thermal domains. In thermal to visual face recognition, a thermal image is a probe image, and a visual image is a gallery image. Thermal to visual face recognition matches thermal probe images with visual gallery images. The availability of training data of thermal and visual image pairs is very difficult, which makes thermal to visual face recognition an even more challenging problem.

In this thesis, the main focus is to learn a generalized thermal to visual face recognition model with less amount of training data. For this, we have proposed three methods. The first method is on deep learning-based transfer learning method, the second is dictionary learning-based common sub-space learning, and the third is collaborative metric learning.

In the first method, we propose a transfer learning approach to enhance the accuracy of the thermal classifier. As a result, the accuracy of thermal-to-visual face recognition is significantly increased. The proposed method is tested on the RGB-D-T dataset (45,900 images) and UND-X1 collection (4,584 images). By transferring knowledge, the experimental results indicate a noticeable increase in the overall accuracy of thermal to visual face recognition.

In the second method, we propose a new common subspace learning approach based on commonality and particularity dictionary learning. In this approach, first, we distinguish the domain-specific and identity-related representations. The domain-specific representation is then removed, and the common subspace is obtained. Finally, the similarity is learnt by the metric learning methods. The proposed method is tested on RGB-D-T and RegDB data sets. The experimental results show that the proposed method performs better even when no common person exists between training and testing sets.

In the third method, we propose collaborative metric learning using maximum margin matrix factorization. This method considers group-wise similarities and collaboratively predicts the similarities. In this method, we can learn a more generalized metric by utilizing the maximized margin. The proposed method is tested on the RGB-D-T and RegDB data sets, and it outperforms the existing works in the few-shot learning settings.

Among the three proposed methods, the two methods (dictionary learning method and collaborative metric learning method) are learnt with less amount of training data. Among these two methods, the collaborative metric learning approach demonstrates superior performance in few-shot learning scenarios.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AE	Auto Encoder
BDTR	Bi-directional Dual-Constrained Top-Ranking
CF	Collaborative Filtering
CML	Collaborative Metric Learning
CNN	Convolutional Neural Network
CSL	Common Subspace Learning
DML	Deep Metric Learning
DSD	Dense Sparse Dense
EER	Equal Error Rate
FIR	Far-Infrared
GAN	Generative Adversarial Network
HOG	Histogram of Oriented Gradients
HTML	Hierarchical Cross-modality Metric Learning
IR	Infrared
LBP	Local Binary Patterns
LSML	Large scale metric learning
LWIR	Long-Wave Infrared
MACE	Modality aware collaborative ensemble learning
mAP	Mean Average Precision
MF	Matrix Factorisation
MMMF	Maximum Margin Matrix Factorisation
MWIR	Mid-Wave Infrared
NIR	Near-Infrared
PLS-DA	Partial Least Squares Discriminant Analysis
ReLU	Rectified Linear Unit
SIFT	Scale-Invariant Feature Transform
SWIR	Short-Wave Infrared
TL	Transfer Learning
TONE	Two-stream CNN network

GLOSSARY

- **Collaborative metric learning:** Collaborative metric learning is a metric learning algorithm. It learns the distance function by capturing the object-level relationships as well as group-level relationships.
- **Common subspace learning:** Common subspace learning is a type of machine learning technique in which multiple datasets are jointly analyzed and a shared subspace is learned that captures the common information or structure among the data.
- **Dictionary learning:** Dictionary learning is a subfield of machine learning in which an algorithm learns to represent a dataset as a linear combination of a set of basis functions or dictionary atoms.
- **Few-shot learning:** Few-shot learning aims to learn a model that can quickly adapt to new tasks with very little data. This is done by training the model on a set of related tasks or classes and then testing it on a new, unseen task or class with just a few training examples.
- **Learning:** When a computer program is able to improve its performance on a particular set of tasks T, as measured by a given performance metric P, through the use of experience E (usually in the form of training data), we say that the program has learned.
- **Metric learning:** Metric learning is a subfield of machine learning that focuses on learning a distance function over the objects. It learns the distance function by capturing the important relationships among objects. The learned function measures the similarity between the objects.
- **One-shot learning:** One-shot learning is a specific case of few-shot learning in which a model is trained to recognize objects or patterns after seeing just a single example of each class. In other words, the model is trained to generalize from a single training example of each class to new, unseen examples.
- **Transfer learning:** Transfer learning is a machine learning technique in which a model trained on one task or domain is leveraged to improve the performance of a related task or domain.
- **Zero-shot learning:** Zero-shot learning is a machine learning approach where a model is trained to recognize objects or patterns without any examples of some classes during training. In other words, the model is trained to generalize to new, unseen classes, which were not present in the training data.

CHAPTER 1

Introduction

1.1 Introduction

In recent years, surveillance systems have become more significant in remote surveillance of people, property, and both public and private sites. The use of surveillance systems has greatly improved security. Because of these applications, surveillance systems have evolved significantly. The face recognition system is a very important part of the surveillance system. The face is one of the most easily accessible biometric modalities that does not require special acquisition procedures and cooperation of the subject, which are reasons that make it useful for a wide variety of biometric applications. Facial recognition technology treats the face as an index of identity. The face recognition system matches the human face image against the existing face image database, and with this matching, it confirms the person's identity. Face recognition has been a very active research topic in computer vision and pattern recognition for decades.

Surveillance in low light or at night is a vital capability of the surveillance system. Surveillance cameras use other spectral bands, such as NIR (near-infrared) and thermal bands, for nighttime surveillance. Infrared (IR) is the most frequently used spectral band in surveillance next to the visual spectrum.

Figure 1.1 illustrates the wavelength range of visual and infrared spectral categories in the electromagnetic spectrum. The infrared spectrum is divided into five categories based on wavelength: Near-Infrared (NIR), Short-Wave Infrared (SWIR), Mid-Wave Infrared (MWIR), Long-Wave Infrared (LWIR) and Far infrared (FIR). The wavelength

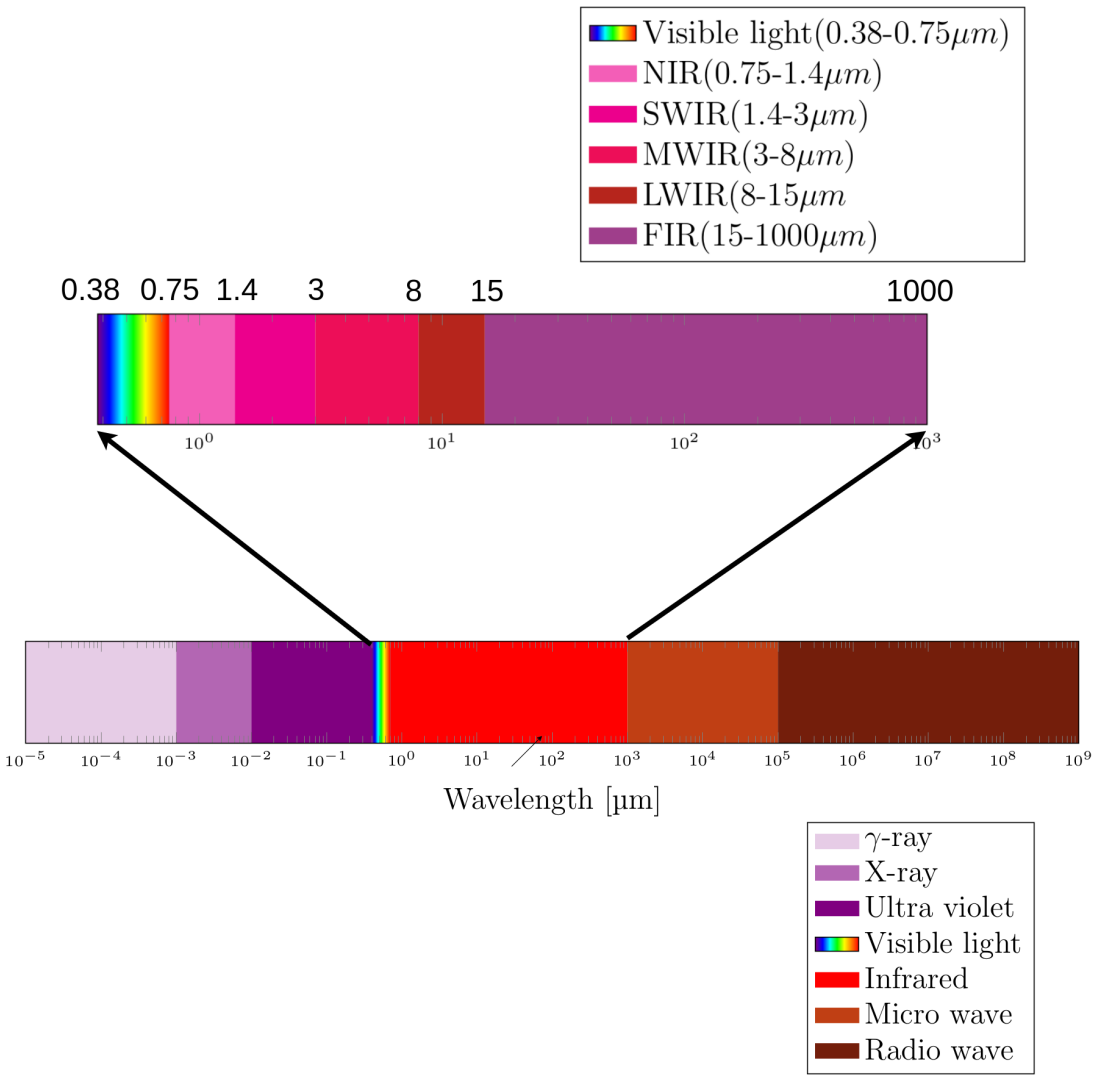


Figure 1.1: Illustration of the wavelength range of Visual and Infrared spectral categories in the electromagnetic spectrum

range of visual and IR categories is shown in Table 1.1.

Spectral Region	Wavelength Range
Visual	$0.38\mu m$ to $0.75\mu m$
NIR	$0.75\mu m$ to $1.4\mu m$
SWIR	$1.4\mu m$ to $3\mu m$
MWIR	$3\mu m$ to $8\mu m$
LWIR	$8\mu m$ to $15\mu m$
FIR	$15\mu m$ to $1000\mu m$

Table 1.1: Visual and infrared spectral regions and wavelength ranges in the electromagnetic spectrum

The reflective-infrared imaging system utilises the NIR and SWIR bands to capture images based on the reflection of infrared light. To facilitate this process, an external infrared source is necessary, which gives these bands a similar illuminating effect as that of the visual spectrum imaging system. Conversely, the thermal-infrared imaging system employs the MWIR and LWIR bands. This system captures images based on the natural thermal emission of objects or bodies. Consequently, the thermal band imaging does not rely on any external illuminating sources.

Thermal imaging has a number of advantages, including the ability to capture images in low light, complete darkness, or other challenging situations such as smoke-filled and dusty environments, as well as the ability to capture images from a large distance. Considering these advantages, the thermal band is widely used in military and security surveillance applications. However, most computer vision research and human visual systems have evolved in the visual spectrum, necessitating thermal to visual face identification.

Figure 1.2 illustrates the consistency of the thermal images with different weather and lighting conditions. Figure 1.2 e is the thermal image. Figure 1.2 a, 1.2 b, 1.2 c, and 1.2 d are the visual images with different weather/light conditions. Here all the images are of the same person.

For automatic surveillance, it is vital to match data gathered at night with data captured during the day, which introduces a new challenge in face recognition which is cross-domain face recognition. Cross-domain face recognition is not only used in automatic surveillance but also utilised in applications of law enforcement. The task of cross-domain face recognition is to actively match sketch to visual, NIR to visual, and thermal to visual face images. At times, cross-domain face recognition is also referred

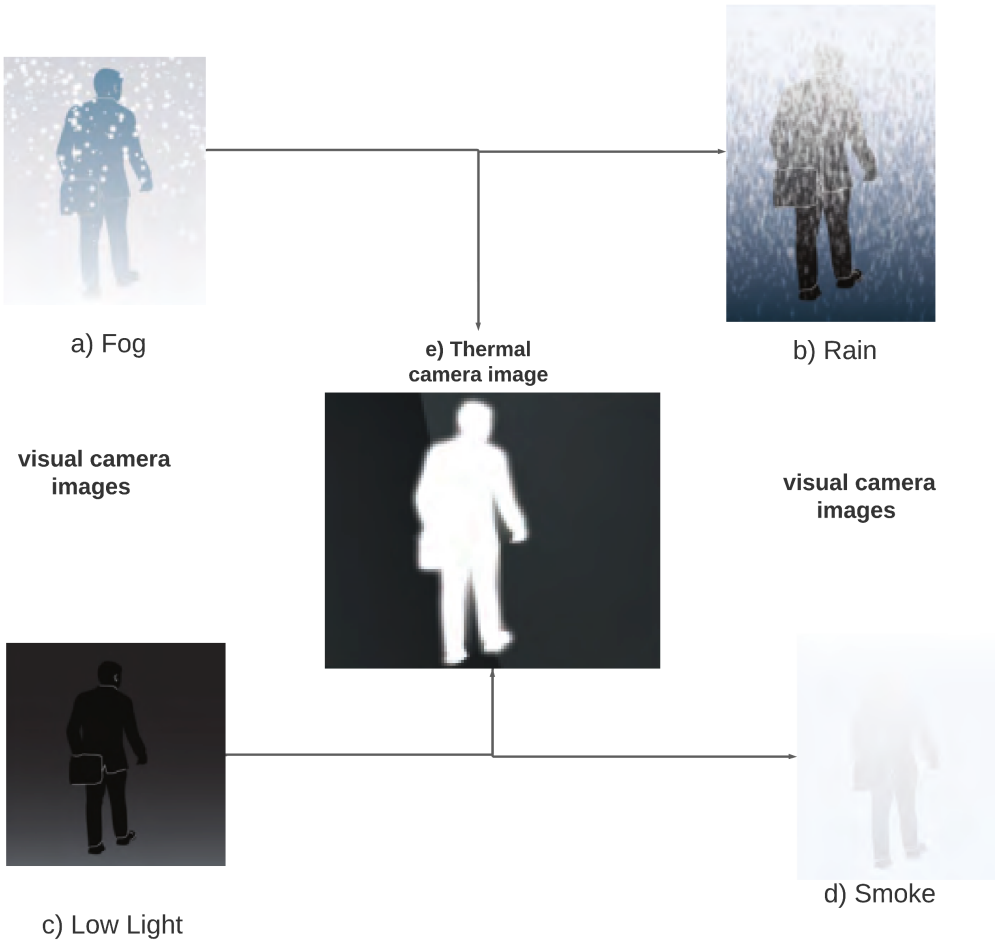


Figure 1.2: Illustration of visual and thermal images for the same person in different weather/light conditions. a), b), c), and d) are visual images with different weather/light conditions. e) is the thermal image for different weather/light conditions.

Figure 1.3: Sample images (a) Visual face image of a person (b) Corresponding thermal domain face image

to as heterogeneous face recognition (HFR). In this work, cross-domain face recognition is done on thermal and visual spectral images. The spectral characteristics of the visual and thermal spectrum are distinct. This results in domain disparity and nonlinear pixel intensity levels between domains. As a result, thermal to visual cross-domain face recognition is more challenging.

The images presented in Figure 1.3 demonstrate the differences in intensity levels between a person's visual face image and its corresponding thermal domain face image. Specifically, the images reveal the non-linear relationship between these domains.

Humans effortlessly identify individuals based on facial features. On the other hand, face recognition algorithms are facing difficulty due to the changes in pose, illumination, expression, occlusion, and disguise. Because of these changes, inter-class and intra-class similarities are increasing. These will adversely affect face recognition accuracy. In addition, domain discrepancies and resolution differences further increase the confusion between inter-class and intra-class similarities.

Figure 1.4 shows pairs of thermal-visual images of the same person, each pair depicting variations in pose, illumination, expression, and occlusion.

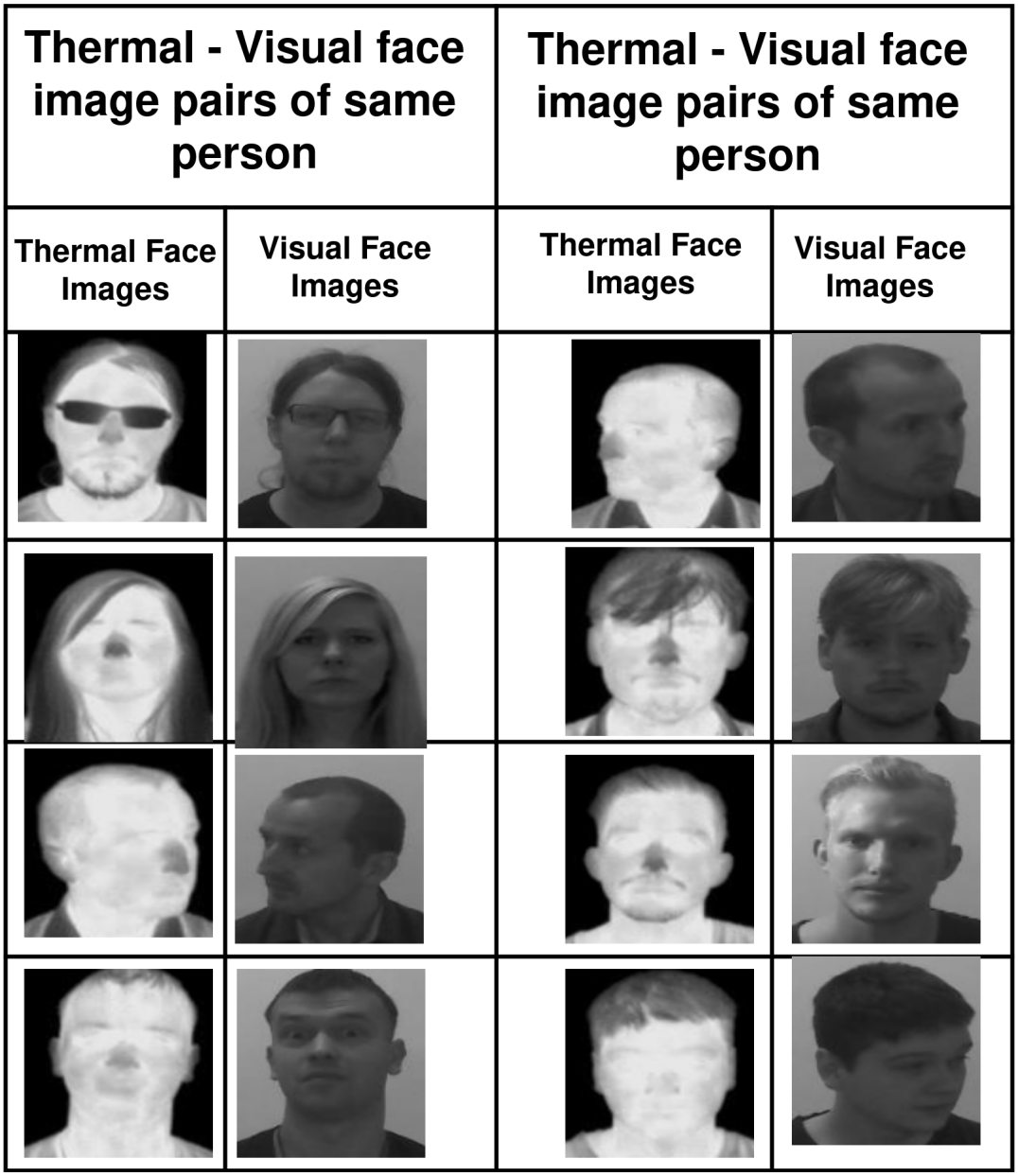


Figure 1.4: Illustration of thermal-visual image pairs of the same person with different pose, illumination, expression, and occlusion

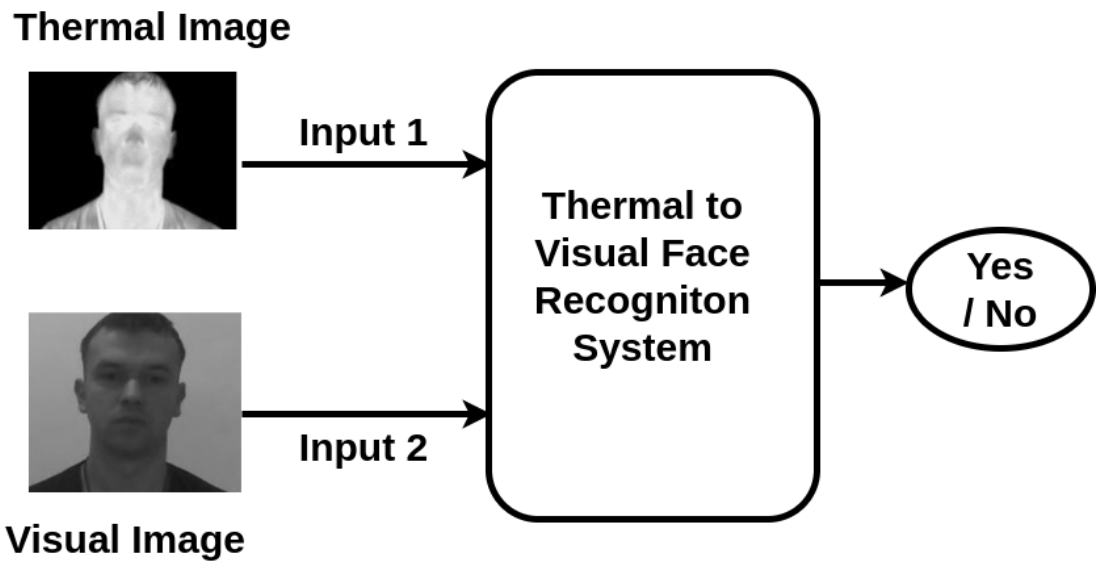


Figure 1.5: Illustration of thermal to visual cross-domain face recognition system

1.2 Problem Statement and Objectives

1.2.1 Problem Statement

Given two input face images, one is from the thermal domain and the other is from the visual domain, the goal is to learn a thermal to visual cross-domain face recognition system which returns 'Yes' if they are of the same person's face; otherwise, it returns 'No'. The same is illustrated in Figure 1.5.

1.2.2 Objectives

- To learn the cross-domain face recognition system by minimising the domain discrepancy.
- To learn a generalised cross-domain face recognition system using fewer training data.

1.3 Contributions

Figure 1.6 illustrates the three distinct contributions of our work, each representing a unique system for thermal-to-visual face recognition. Firstly, a deep learning based transfer learning technique is introduced. Secondly, a common subspace tech-

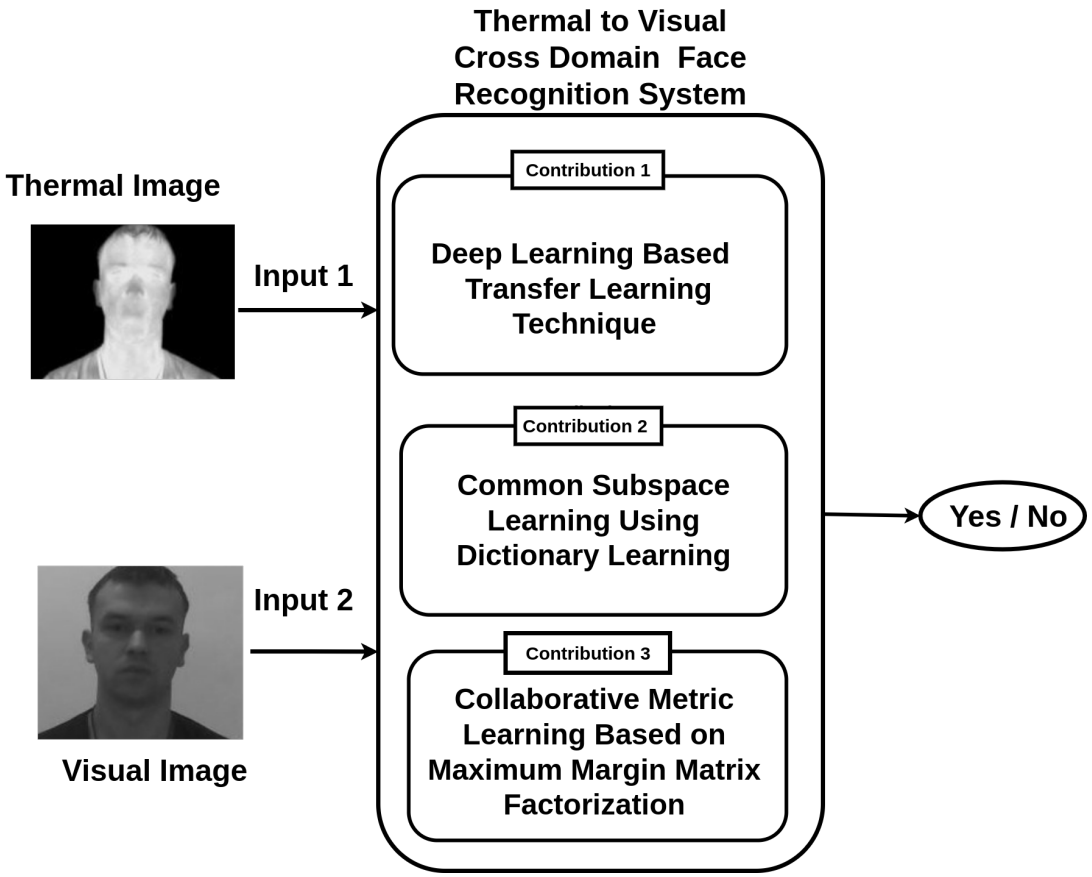


Figure 1.6: Illustration of contributions

nique based on dictionary learning is presented. Lastly, a collaborative metric learning technique employing matrix factorisation is proposed.

1.3.1 Deep learning based transfer learning technique

In this approach, the training of the thermal classifier is accomplished by utilising the visual classifier through transfer learning. Traditionally, transfer learning in deep neural networks involves transferring the weights [102, 163]. However, in this method, we initially sparsify the network before transferring the weights of the sparsified network. To obtain the sparsified network, we propose two pruning methods. By using this method, the knowledge from the visual classifier gets effectively transferred to learn the thermal classifier, and as a result, this thermal classifier gets trained with even less amount of training data.

1.3.2 Common subspace learning using dictionary learning

In this work, we introduce a two-stage cross-domain (thermal to visual) face recognition method based on dictionary learning. We begin by projecting both domain images onto a common subspace, where the face images are represented by a representation code. Metric learning is used in the second stage to measure the degree of similarity between corresponding representation codes. We use commonality and particularity dictionary learning to find the common subspace. In the second stage, we used two variants of metric learning methods one is large scale metric learning method, and the other is deep metric learning method.

1.3.3 Collaborative metric learning based on maximum margin matrix factorisation

In this work, we have proposed a collaborative metric learning method. We find a latent space for this metric using maximum margin matrix factorisation by preserving the training similarities. This latent space is learned collaboratively. On the other hand, image space to the learned latent space mapping is done using a convolution neural network. Using the mapping function, the predicted latent space representation measures the similarity between the images.

1.4 Thesis Outline

This thesis is divided into six chapters, beginning with an introductory chapter and ending with a conclusion chapter. Chapter 2 is devoted to a literature review on cross-domain face recognition. Our three contributions are discussed in Chapters 3 to 5. The outline of the thesis is shown in Figure 1.7.

- **Chapter 1:** ***Introduction***
  Cross-domain face recognition is introduced in Chapter 1. It also provides the motivation and necessary background for the work presented in this thesis. It concludes by providing the scope of the thesis.
- **Chapter 2:** ***Literature Survey***
  Chapter 2 provides an overview of the state-of-the-art methodologies in cross-domain face recognition. We have categorised them into different types of cross-

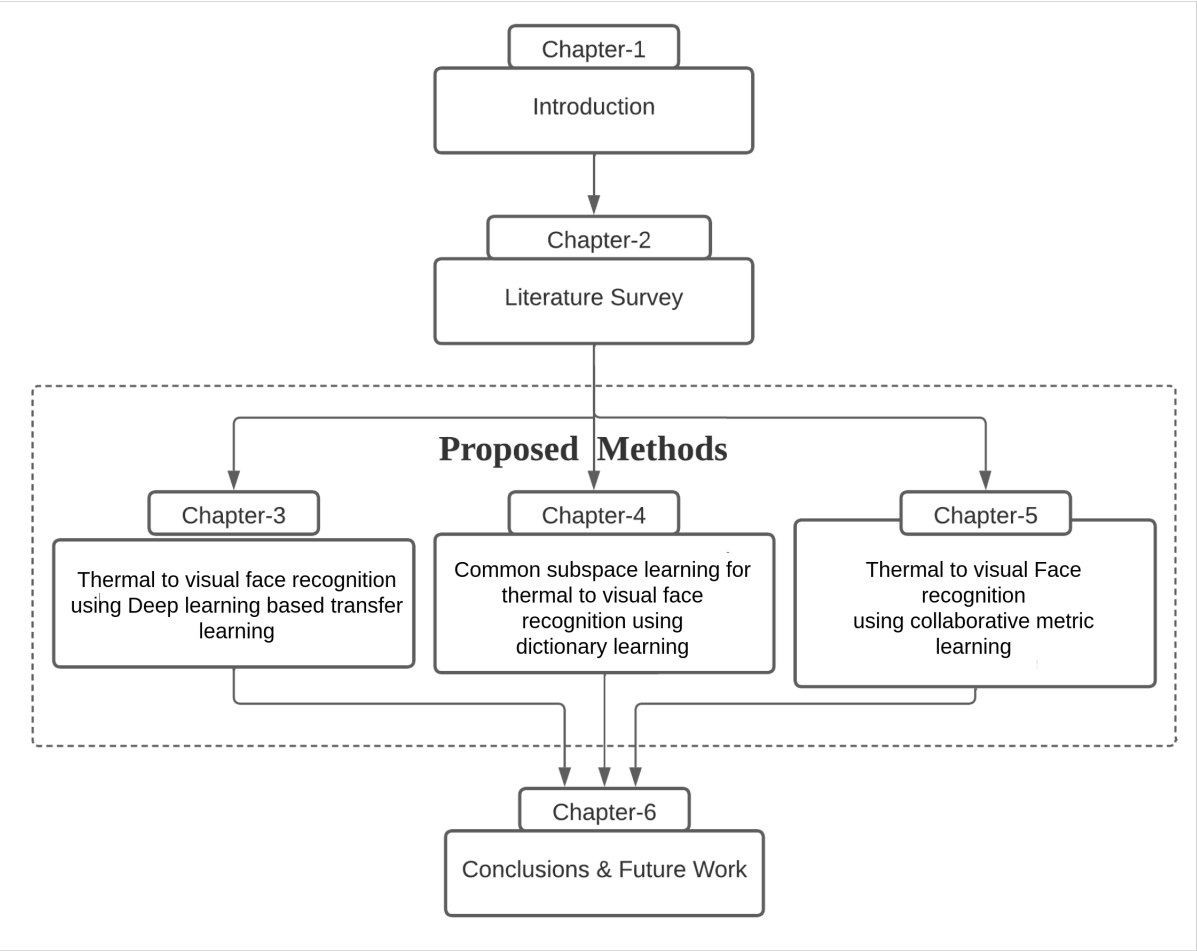


Figure 1.7: Thesis Outline

domain face recognition systems and discussed the scope of the current work in the literature.

- **Chapter 3:** ***Thermal to visual face recognition using Deep learning based transfer learning***
  In this chapter, we discuss the proposed deep transfer learning method. In this approach, the training of the thermal classifier is accomplished by utilising the visual classifier through transfer learning. Traditionally, transfer learning in deep neural networks involves transferring the weights [102, 163]. However, in this method, we initially sparsify the network before transferring the weights of the sparsified network. To obtain the sparsified network, we propose two pruning methods. By using this method, the knowledge from the visual classifier gets effectively transferred to learn the thermal classifier, and as a result, this thermal classifier gets trained with even less amount of training data.

- **Chapter 4:** ***Common subspace learning for thermal to visual face recognition using dictionary learning***
  This chapter introduces the proposed common subspace learning method using dictionary learning. In this work, we have proposed a two-stage cross-domain (thermal to visual) face recognition method based on dictionary learning. We begin by projecting both domain images onto a common subspace, where the face images are represented by a representation code. Metric learning is used in the second stage to measure the degree of similarity between corresponding representation codes. We use commonality and particularity dictionary learning to find the common subspace. In the second stage, we used two variants of metric learning methods one is large scale metric learning method, and the other is deep metric learning method.

- **Chapter 5 :** ***Thermal to visual Face recognition using collaborative metric learning***
  In this chapter, we discuss the proposed collaborative metric learning method. We find a latent space for this metric using maximum margin matrix factorisation by preserving the training similarities. This latent space is learned collaboratively. On the other hand, image space to the learned latent space mapping is done using a convolution neural network. Using the mapping function, predicted latent space representation measures the similarity between the images.

- **Chapter 6 :** ***Conclusions and Future Directions***
  We conclude our thesis with Chapter 6 by giving future directions for extending our work.

CHAPTER 2

Literature Survey

In this chapter, we first review the related work before going through a literature survey on thermal to visual cross-domain face recognition. And then, we discuss the datasets and metrics used for thermal to visual cross-domain face recognition. In the end, we discuss the challenges of thermal to visual cross-domain face recognition.

2.1 Related Work

In many machine learning algorithms, the assumption is that the training and future data are in the same feature space and distribution. Here, future data refers to the unseen data that a trained machine learning model encounters and needs to make predictions on after the training phase. But for many real-world applications, this assumption may not hold, and as a result, the performance of those may affect[147]. Face recognition is one of that applications[5]. Automatic face recognition is a well-explored problem in computer vision and machine learning and performs well only in controlled settings[60]. Face recognition in an uncontrolled environment is still a challenging problem[26, 39]. It is challenging[89] because of different illumination conditions[14, 75],pose[27, 170], camera view angles[100, 169], age-related facial changes[121, 137], facial expressions[67, 107], etc.

In addition to these challenges, many face recognition applications use other than the visual domains. For example, nighttime security applications[62] capture face images in NIR (near infrared)[29] or thermal domain[71]. Some law enforcement appli-

cations use sketch images[56]. Considering these applications, there is an increase in the usage of domains other than the visual domain. But the human visual system and most computer vision research are evolved in the visual domain. As a result, there is a necessity to match visual domain images to other domain images. This requirement is evolving as the new research direction is cross-domain face recognition.

Cross-domain face recognition is a multifaceted task that aims to accurately match individuals across diverse domains[103]. These domains encompass various scenarios such as Near-Infrared (NIR) to visual face recognition[46, 47], sketch to visual face recognition[54, 148], high-resolution to low-resolution face recognition[95], 3D image to 2D image face recognition[108], and thermal to visual face recognition[23, 158]. This task becomes even more challenging due to the variations in lighting conditions, camera viewpoints, ages, and ethnicities encountered in different domains. Successfully addressing cross-domain face recognition requires robust algorithms capable of effectively handling the significant variations in facial appearance arising from these diverse factors and modalities.

Thermal to visual cross-domain face recognition[23] is a specialised subfield of cross-domain face recognition that focuses on recognising faces across different modalities, specifically the thermal and visual spectrums. Thermal imaging captures the thermal emission of objects, including a person's face. These distinctive thermal patterns are used for face recognition. On the other hand, visual imaging captures the appearance of faces in the visible light spectrum.

Thermal imaging[71] has been widely used in various applications such as surveillance, law enforcement, and night-time monitoring, where visual imaging may be limited due to lighting conditions or environmental factors. However, there are significant challenges in recognising faces across thermal and visual modalities, as the two modalities have different imaging properties, such as pixel intensities, texture patterns, and illumination characteristics. The domain differences between thermal and visual modalities pose a major obstacle for conventional face recognition systems. Models trained on visual images may not excel when dealing with thermal images, and vice versa.

In the literature, most of these challenges are addressed with transfer learning[104, 147] by adequate knowledge transferring. Various transfer learning approaches are pro-

posed for thermal to visual cross-domain face recognition techniques. These techniques aim to leverage knowledge from one domain (e.g., visual) to improve face recognition performance in another domain (e.g., thermal) by developing methods to mitigate the domain discrepancy and learn discriminative features invariant to the domain.

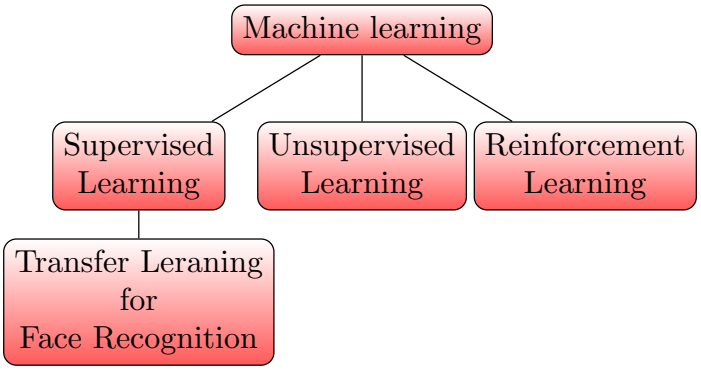


Figure 2.1: Categories of machine learning

Figure 2.1 illustrates the three main categories of machine learning[10]: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the algorithm is provided with labelled data to learn a function that maps input features to output labels. In unsupervised learning, the algorithm discovers patterns or structures in the input data without any corresponding output labels. In reinforcement learning, the algorithm learns by interacting with an environment and receiving feedback through rewards or penalties. Transfer learning comes under supervised learning.

Figure 2.2 illustrates the taxonomy based on our literature survey. This taxonomy provides a comprehensive overview of transfer learning for face recognition. Here, red nodes indicate related work, while blue nodes represent literature survey on cross-domain face recognition.

After conducting an extensive literature survey, we have constructed the taxonomy. The survey encompassed approximately 170 papers and 7 survey papers, covering a wide range of topics, including covariate shift problems [14, 27, 78], multimodal face recognition [176], multi-task learning [167], and heterogeneous face recognition [103, 140]. Our exclusive focus was on thermal to visual cross-domain face recognition, and we reviewed all the relevant papers published between 2012 and 2020.

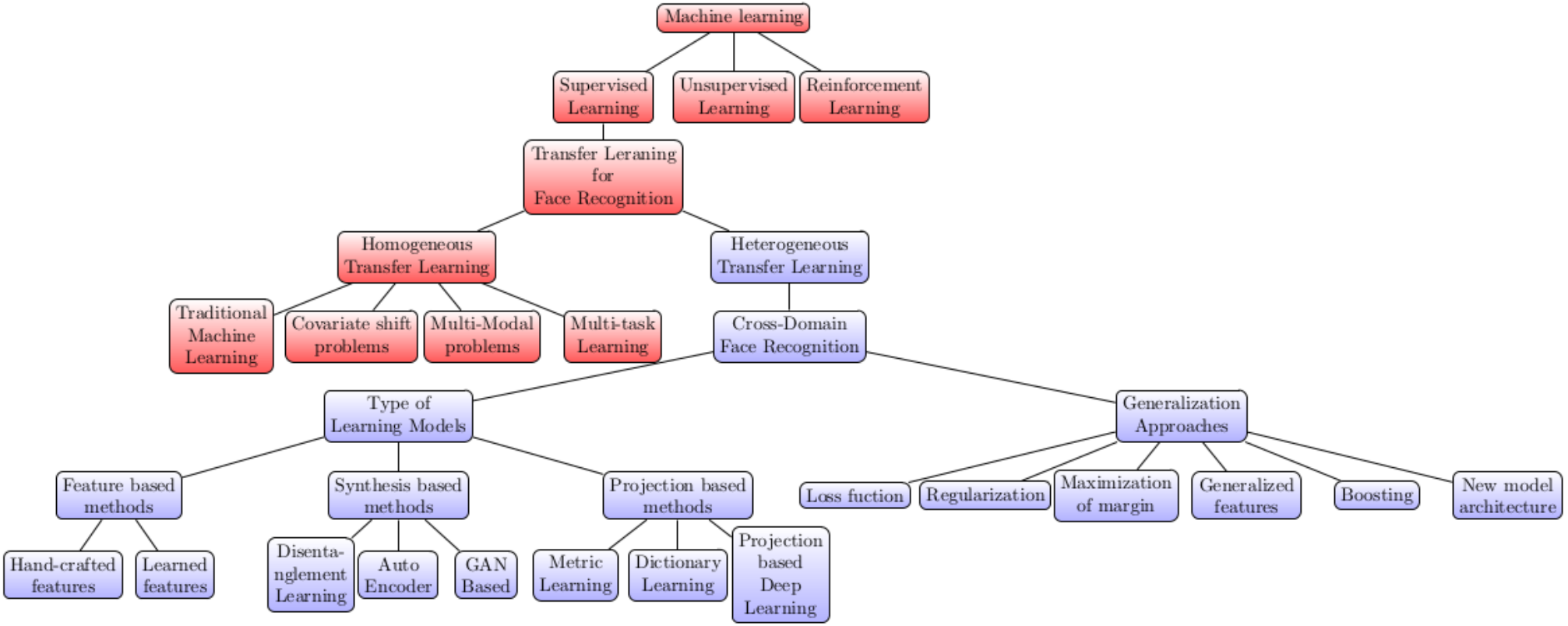


Figure 2.2: Taxonomy of transfer learning for face recognition: red nodes indicate related work, while blue nodes represent literature survey on cross-domain face recognition.

2.2 Transfer learning for face recognition

Transferring knowledge across tasks is a natural ability for humans, i.e., knowledge acquired while learning about one task is reused while solving other related tasks. For example, knowledge gained while riding a motorbike (task) can be used to learn the driving of a car (related but different task). Transfer learning [104][147] is formally defined as follows. A domain $\mathcal{D}$ contains two parts - a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ (as we consider a subset of random variables, the marginal probability is used). Here X is a given sample set $X = \{x_1, ..., x_n\}$ in the feature space $\mathcal{X}$, with corresponding labels set Y in label space $\mathcal{Y}$. Task T is defined using two parts, a label space $\mathcal{Y}$ and a predictive function $h(.)$. In general $h(.)$ is a conditional probability $P(Y/X)$. Let the source domain be $\mathcal{D}^s$, where $\mathcal{D}^s = \{\mathcal{X}^s, P(X^s)\}$, and the corresponding source task is T^s, where $T^s = \{\mathcal{Y}^s, h^s(.)\}$. Let the target domain be $\mathcal{D}^t$, where $\mathcal{D}^t = \{\mathcal{X}^t, P(X^t)\}$, and the corresponding target task is T^t, where $T^t = \{\mathcal{Y}^t, h^t(.)\}$. Now the goal of transfer learning is to improve the target predictive function $h^t(.)$ by using the related information from source domain $\mathcal{D}^s$ and source task T^s.

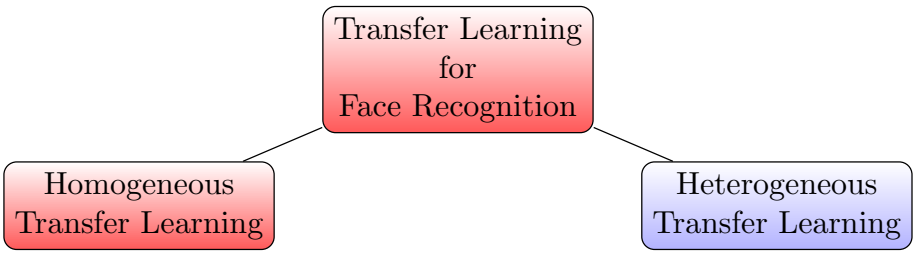


Figure 2.3: Categories of transfer learning for face recognition

Figure 2.3 shows the two types of transfer learning: homogeneous transfer learning and heterogeneous transfer learning.

Transfer learning is divided into two categories based on feature space $\mathcal{X}$.

- **Homogeneous transfer learning** In homogeneous transfer learning, source domain feature space and target domain feature space is the same ($\mathcal{X}^s = \mathcal{X}^t$).
- **Heterogeneous transfer learning** In heterogeneous transfer learning, source domain feature space and target domain feature space is different ($\mathcal{X}^s \neq \mathcal{X}^t$).

2.2.1 Homogeneous transfer learning

Homogeneous transfer learning refers to the application of transfer learning techniques within the same or similar domains. Homogeneous transfer learning has four major categories.

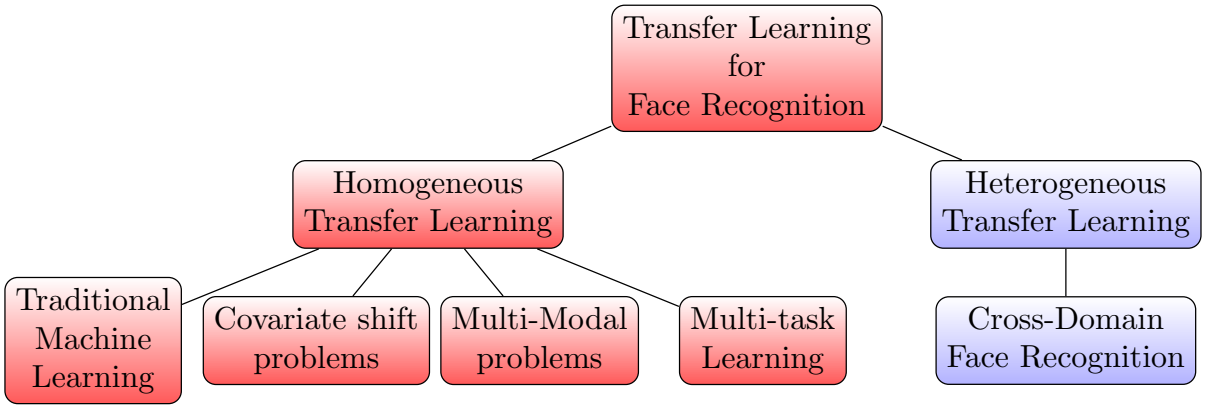


Figure 2.4: Categories of homogeneous transfer learning for face recognition

Figure 2.4 shows the four categories of homogeneous transfer learning. They are

- Traditional machine learning.
- Covariate shift problem.
- Multimodal face recognition.
- Multi-task learning.

2.2.1.1 Traditional machine learning

Traditional machine learning problems have the same domain and the same task. Same domain means having same feature space ($\mathcal{X}^s = \mathcal{X}^t$) and same marginal distribution ($P(X^s) = P(X^t)$). Same task means, having same label space ($\mathcal{Y}^s = \mathcal{Y}^t$) and same predictive function ($h^s(.) = h^t(.)$). So, for traditional machine learning algorithms, the requirement of knowledge transfer between source and target is limited. Face recognition in a controlled environment [133, 172] is an example problem that falls under this category.

2.2.1.2 Covariate shift problem

In homogeneous transfer learning problems, if the task is the same, but the marginal probability distribution is different for the source and target, then it is a covariate shift problem. Covariate shift problems are also called as dataset bias problems. In this category of problems, the training set and test set are in different marginal probability distributions.

Some of the examples of covariate shift problems are person re-identification in surveillance[123, 165], makeup invariant face recognition [38, 143], and face recognition after plastic surgery [86, 128]. In all of these problems train set is in one probability distribution, and the test set is in a slightly different marginal probability distribution.

2.2.1.3 Multi-modal face recognition

For traditional machine learning and multimodal face recognition, feature space and marginal probability are the same for both source and target. But the main difference is the formation of feature space, and the probability distribution is the union of different modalities. In multi-modal data, $X = \{X_1, X_2, ..., X_m\}$, X_i refers to m number of different modalities. Each modality has n different images of a particular modality, and feature space is the union of all modalities $\mathcal{X} = \{\mathcal{X}^1 \cup \mathcal{X}^2 \cup, ..., \cup \mathcal{X}^m\}$. The marginal probability distribution of multimodal face recognition is the joint distribution of other modalities, $P(X) = P(X^1, X^2, ..., X^m)$. The main challenge of multimodal face recognition is in coordinating different modalities.

Multimodal face recognition involves recognising individuals using multiple biometric data sources, such as images from different modalities. Multimodal face recognition is considered more robust than single-modal approaches. It has been applied in various domains, such as visual+3D [15, 58], visual+IR [12, 21], visual+thermal [4, 9] and visual+IR+3D [11].

2.2.1.4 Multi-task learning

Multi-task learning is the task of learning multiple tasks at a time. Here both domains are same $\mathcal{D}^s = \mathcal{D}^t$ and each task has its own label space $\mathcal{Y}_1, \mathcal{Y}_2, ...\mathcal{Y}_n$. For

multi-task learning, the predictive function is $P(\mathcal{Y}_1, \mathcal{Y}_2, ...\mathcal{Y}_n/X)$.

Multi-task learning for face recognition involves training a model to simultaneously perform multiple tasks related to face recognition. Examples of multi-task learning problems are face recognition along with expression [91], face recognition along with age [65], face recognition along with gender [6], face recognition along with ethnicity [25], face recognition along with landmark localisation [171].

2.2.2 Heterogeneous transfer learning

Heterogeneous transfer learning involves scenarios where the feature spaces in the source and target domains differ, while the tasks remain the same. In heterogeneous transfer learning, the feature spaces are different, $\mathcal{X}^s \neq \mathcal{X}^t$, and the tasks are the same $T^s = T^t$. In heterogeneous transfer learning, only one valid category is cross-domain face recognition.

2.2.2.1 Cross-domain face recognition

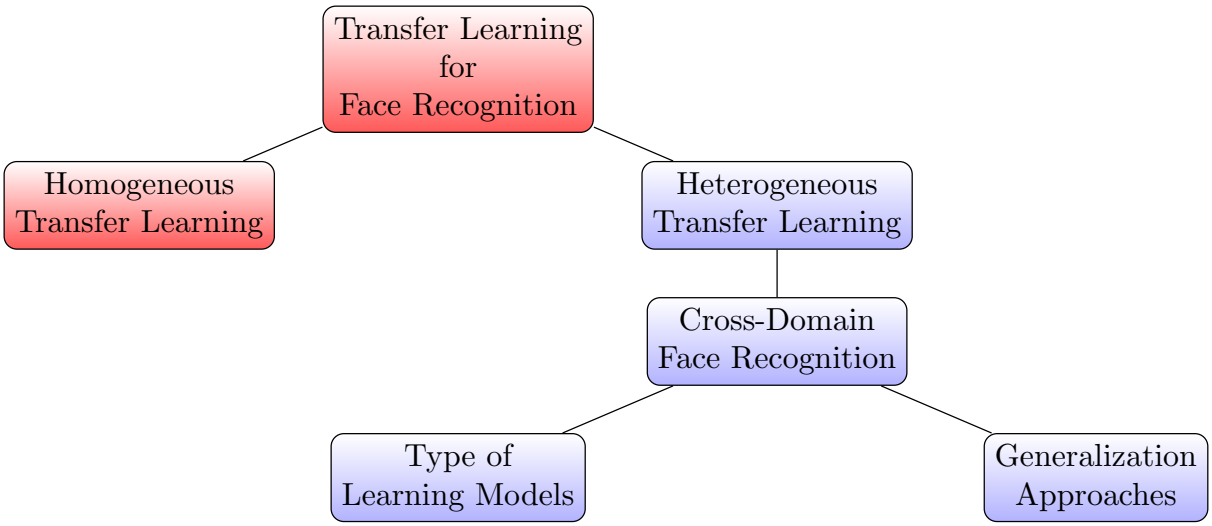


Figure 2.5: Categories of cross-domain face recognition

Cross-domain face recognition comes under heterogeneous transfer learning [103]. Cross-domain face recognition is a challenging task because of domain discrepancy. At times, cross-domain face recognition is also referred to as heterogeneous face recognition, cross-domain person re-identification and inter-modality face recognition. The

literature studies cross-domain face recognition on NIR-visual[46, 47], sketch to visual [54, 148], 3D-2D images[108], low-resolution - high-resolution[95], and thermal-visual[23, 158]. Early research on cross-domain face recognition predominantly concentrated on the sketch to visual and NIR to visual. Cross-domain face recognition literature is categorised in two different ways.

Figure 2.5 illustrate the cross domain face recognition literature categorisation approaches. They are

- Type of learning model.
- Generalisation approaches.

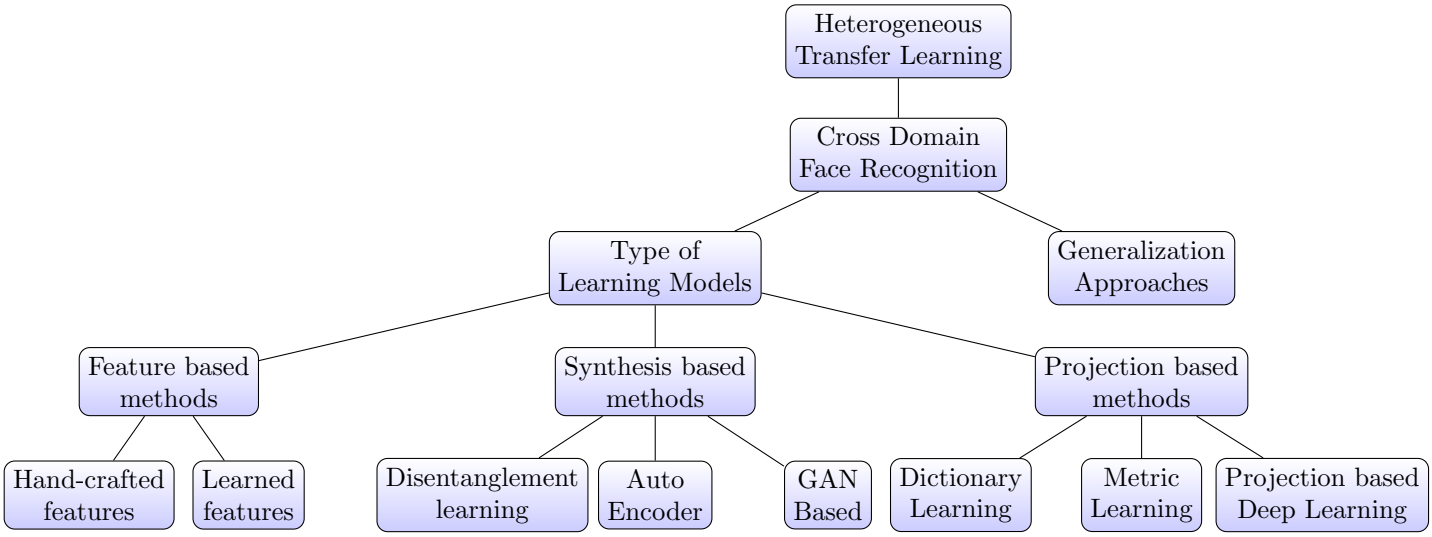


Figure 2.6: Cross-domain face recognition literature categories based on the type of learning model

2.2.2.1.1 Type of learning models: Figure 2.6 illustrates the categories and subcategories (based on the literature) of cross-domain face recognition based on the type of learning model. Cross-domain face recognition is categorised into three [103], namely

a) Feature-based methods

b) Synthesis-based methods

c) Projection-based methods

Table 2.1 summarises the categories of cross-domain face recognition literature into three categories - feature-based, synthesis based and projection based.

Feature Based methods	Hand-Crafted Features	SIFT, HOG, and LBP are used in [23, 51, 120] 2D-LDA used in [168], Logarithm gradient histogram(LGH) is used in [178], Local Quality Descriptor is used in [8]
	Learned Features	Discriminative features are learning in [83, 92, 151, 157, 158], Hierarchical boosting network [152]. Single hidden-layer Gabor-based network [101]. Hard modality alignment network (HMAN) for modality-robust features [141]. CNN-based features [97, 149]. Feature learning method based on iterative closest point method [131]. New feature mapping sub-network [32]. HFR framework for matching visual and thermal face images [17]. Relational graph module (RGM) [22]. Two-stream network with part-level person feature learning [84].
Synthesis Based Methods	Disentanglement learning	GAN based disentanglement learning methods are proposed in [53, 63, 68, 105, 156]
	Auto Encoder	Auto encoder based synthesis model is proposed in [64, 126]
	GAN	GAN based synthesis models are proposed in [18, 44, 47, 63, 66, 68, 106, 145, 166]
Projection Based Methods	Dictionary Learning	Dictionary based methods are proposed in [88, 94, 96, 114, 115]
	Metric Learning	Metric learning based methods are proposed in [46, 59, 82, 94, 117, 158, 161, 162]
	Projection based Deep Learning	Deep learning based methods are proposed in [34, 74, 79, 81, 175]

Table 2.1: Cross-domain face recognition literature categories based on the type of learning model

a) **Feature based methods:** Feature-based techniques [37] aim to choose and extract domain-invariant features. The feature-based method is further divided into two categories.
   - Hand-crafted features
   - Learned features
   - **Hand-crafted features:** The features which are extracted manually using a predefined algorithm are called hand-crafted features. In [23, 51], the authors introduced a partial least square discriminate analysis method (PLS-DA) that matches the extracted handmade features like SIFT, HOG, and LBP from both visual and thermal domain images. In [120], hand-crafted features are extracted and matching with a deep matching algorithm is done. In [168], low-resolution images with high-resolution (HR) images based on two-dimensional linear discriminant analysis (2D-LDA) are performed. In [178], a logarithm gradient histogram (LGH) is proposed for the NIR to visual cross-domain face recognition. In [8], the authors proposed a local quality descriptor. In [119], local mesh pattern (DLMeP) is derived and which is used to measure the local variation or pattern of wavelet energy.
   - **Learned features:** The features learned by the trained model are obtained automatically, without manual intervention [151]. In order to enhance the discriminative power of the learned features, various approaches have been proposed in the literature.

     One such approach is hierarchical discriminant feature learning (HDFL), which was introduced by Xu et al. [151]. Another method proposed by Xu et al. [152], involves using a hierarchical boosting network. Oh et al. [101] proposed a single hidden-layer Gabor-based network for feature learning.

     To address modality discrepancies in features, Wang et al. [141] proposed the hard modality alignment network (HMAN), which aims to learn modality-robust features. Additionally, CNN-based features have been utilised in feature learning, as demonstrated in the works of Wu et al. [149] and Nguyen et al. [97].

     In the thermal imaging domain, Sun et al. [131] proposed a feature learning method based on the iterative closest point method. Nimpa et al. [32] introduced a new feature mapping sub-network to improve performance. Hu et al. [52] proposed sparse multiple kernel learning (SMKL) for feature extraction in thermal images.

     For matching visual and thermal face images, Chen et al. [17] developed the high-frequency representation (HFR) framework, which matches images using multiple subspaces generated from patches.

     To incorporate relational information into the feature learning process, Cho et al. [22] introduced the relational graph module (RGM), a graph-structured module.

     In the context of modality discrepancies, Ye et al. [157] introduced the MACE learning method, which focuses on learning discriminative middle-level features and addresses the differences between modalities in features and classifiers [157, 158].

     Liu et al. proposed two methods for feature learning. In [84], a two-stream network with part-level person feature learning was proposed. In [83], an enhanced discriminative feature learning method was introduced [83, 84].

b) **Synthesis-based methods:** In synthesis-based methods [109, 139], the main focus is on synthesising target domain images to source domain images. Synthesis methods mainly use three types of architectures, namely
   - Disentanglement learning
   - Auto encoder based (AE)
   - Generative adversarial network (GAN)
   - **Disentanglement learning:** A disentangled representation is one in which the changes in single latent units are sensitive to changes in one generative factor but are not affected by changes in other factors. GAN-based disentangle representation is proposed in [53, 63, 68, 105, 156]. These methods try to separate the identity-related features and perform cross-domain face recognition.
   - **Auto encoder based (AE):** Auto encoder based methods have two parts - encoder and decoder. The encoder tries to convert the target domain image to the source domain image, and the decoder tries to convert the source domain image to the target domain image. In [126], an auto-encoder-decoder-based network is proposed for cross-domain face recognition. In [64], deep autoencoder based method is proposed
   - **Generative adversarial network (GAN):** The generative adversarial network (GAN) is a technique that is commonly utilised in synthesis-based approaches [145, 166]. Using GAN, target domain images are synthesised into the source domain. GAN-based bidirectional heterogeneous prototype learning is proposed in [106]. In [47], the authors proposed GAN based end-to-end NIR-VIS face completion network. In [44], Modality Adversarial Neural Network (MANN) is proposed to extract modality-invariant features. In [63, 68], stack of GANs are used for the synthesis. In [66], thermal to rgb generative adversarial network (TRGAN) is proposed. In [18], semantic-guided generative adversarial network (SG-GAN) is proposed.

c) **Projection based methods:** Projection-based methods are a widely used approach for the cross-domain face recognition problem. In this approach, both the domain images are projected into a shared subspace using projection-based techniques [116], which are more comparable than in the original space. Projection-based methods are further divided into three categories.
   - Dictionary learning based methods
   - Metric learning methods
   - Projection-based deep learning methods
   - **Dictionary learning based methods:** In dictionary learning, the given data is split into two parts - one is the dictionary, and the other is the corresponding representation code. The dictionary is a set of atoms, and these atoms are basis vectors for new projected subspace. In [114], coupled dictionary learning method is proposed. In this method, for each domain, one dictionary is learned. In [96], dictionary learning-based discriminative shared transform learning (DSTL) is proposed. In [88], semi-coupled mapping and discriminant dictionary learning (SMD2L) is proposed. In [94], the authors proposed dictionary learning based on common subspace learning. In [115], sparse representation is learned based on random subspace base learners.

– **Metric learning methods:** Metric learning is another projection-based method. In metric learning, both the domain images are projected in a common subspace and then in the new space estimate the similarity. In [46], the Wasserstein distance-based layer is introduced. In [117], the authors proposed a neural network-based coupled architecture network, which forces the hidden layers of two neural networks to be as similar as possible. In [94], a Dictionary learning-based metric learning approach is proposed. In [59] proposed a deep metric learning-based method using maximum mean discrepancy-based loss. In [161], the authors proposed a deep metric learning-based method using intra-modality weighted-part aggregation loss. In [158], the authors proposed a deep metric learning-based method that uses a joint loss of verification and ID loss. In [82], a novel framework for VT-REID addresses the cross-modality gap and intra-modal variations. It incorporates class-aware modality mix for pixel-level gap reduction and center guided metric learning for inter- and intra-modal discrepancy reduction. In [138], two-stage metric learning (TML) method is proposed. It uses local and global metric learning successively.

– **Projection-based deep learning methods:** Deep learning models project both the domain images into a common subspace. In [34], 3-D morphable model is (acts as a common subspace) used to synthesise photographs and sketches. In [79], the authors proposed a sparse coupled projection method using multidimensional scaling joint L2,1-norm regularisation (MDSL21). In [175], the deep coupled spectral regression-based method is proposed. In [74], domain-based angular margin loss and a maximum angular loss are proposed. In [81], shared discriminative feature representation is learned. Dual-path local information structure (DLIS) with position attention-guided learning module (PALM) is proposed In [149]

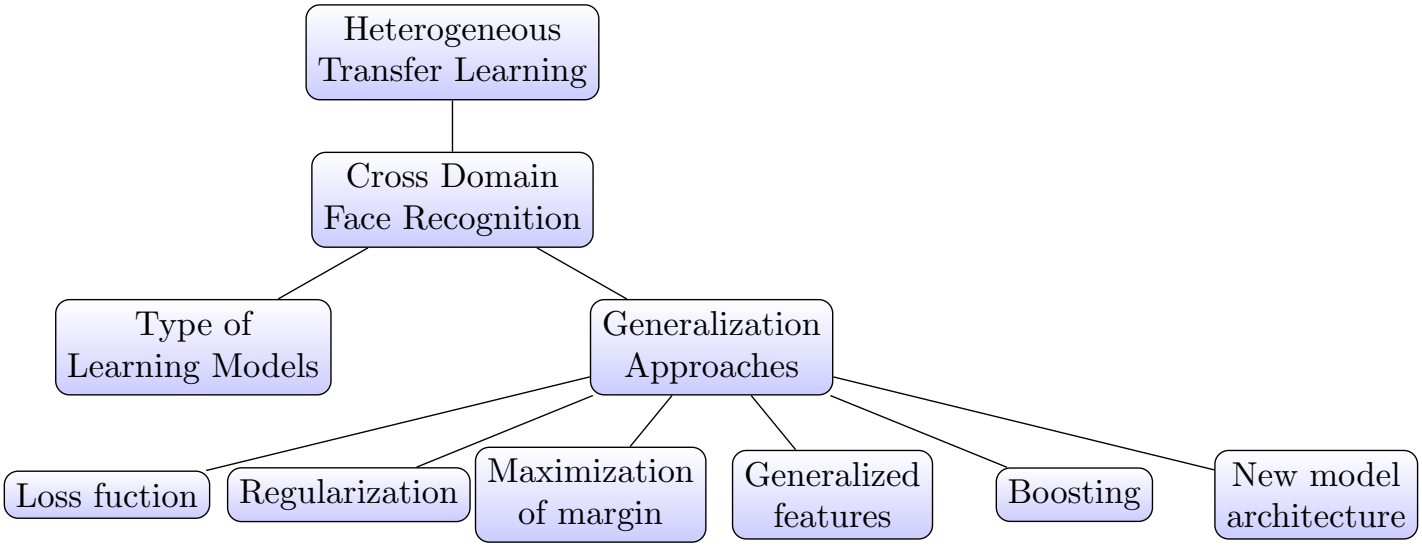


Figure 2.7: Cross-domain face recognition literature categories based on generalisation approaches

2.2.2.1.2 Generalisation approaches: Figure 2.7 illustrates the categories of the literature on cross-domain face recognition based on generalisation approaches. The main objective of transfer learning is to acquire a greater amount of transferable knowledge. This transferable knowledge can be obtained by increasing the level of generalisation in the model [3]. As the model becomes more generalised, the amount of transferable knowledge increases.

Generalisation is a common phenomenon in human and animal learning, where prior knowledge is utilised in various situations and circumstances. It involves the utilisation of prior learning for various situations in different circumstances. Prior learning allows the acquisition of knowledge, which can then be applied to solve future unknown situations in unknown circumstances, making that knowledge transferable. Generalisation plays a vital role in both quantitative and qualitative research [30, 110]. It is particularly significant in quantitative research, especially in pattern recognition and machine learning, where it is one of the core goals [10]. Generalisation is closely tied to the transfer of knowledge. Increased generalisation implies a greater transfer of knowledge. In the context of pattern recognition and machine learning, generalisation refers to the ability to correctly categorise or predict unseen data, which is data not present in the training set.

Over-fitting is a common issue in machine learning that occurs when the model lacks generalisation. It leads to good performance on the training data but poor performance on the test data. To avoid over-fitting, machine learning algorithms generalise the model by acquiring transferable knowledge.

Mainly six categories in cross domain face recognition based on generalisation approches. They are

a) Loss function

b) Regularisation

c) Maximisation of margin

d) Generalised features

e) Usage of Boosting

f) New model architecture

a) **Loss function:** Loss function measures the discrepancy between the observed entry and the corresponding prediction. The loss function is a very important tool for generalisation. The choice of loss function influences many factors like the type of task, data distribution and degree of noise in the data. In [47], 3D-based pose correction loss, two adversarial losses and a pixel loss are proposed. HR-anchor loss is proposed in [127]. In [125] model is jointly trained on two proposed losses: (i) Derived-margin softmax loss and (ii) Reconstruction-center (ReCent) loss. In [33], pairwise identity preserving loss is proposed. In [74], domain-based angular margin loss and a maximum angular loss are introduced. In [45], hypersphere manifold embedding is learned using Sphere Softmax loss. In [44], dual-constrained triplet loss is introduced. In [142], multi patch modality alignment (MPMA) loss is proposed. In [18], semantic loss function is introduced to regularise the adversarial network. In [141], hard modality alignment (HMA) loss is proposed. In [138], mixed-modality triplet loss is proposed.

b) **Regularization:** Regularization is one of the important approaches for generalisation. It avoids overfitting, and it improves generalisation. In [79], a sparse coupled projection method is proposed using multidimensional scaling joint L2,1-norm regularisation.

c) **Maximisation of margin:** In machine learning, especially for the classification tasks maximising the margin is directly proportional to maximising the generalisation. Here maximising the margin means maximising the margin between the classes. In [74], the authors proposed angular margin loss, which maximises the margin. In [125], Derived-Margin softmax loss is proposed to maximise the margin. In [148], coupled deep learning (CDL) method is proposed with an objective function comprising trace norm, block-diagonal prior, and cross-modal ranking to maximise identity margin.

d) **Generalised features:** Feature learning is very crucial in machine learning methods. The generalisation of the model directly depends on the generalisation of the features. SIFT, HOG, and LBP features are used in [23, 51, 120]. 2D-LDA features are used in [168]. Logarithm gradient histogram (LGH) features are used in [178], Local Quality Descriptor features are used in [8], Hierarchical discriminant feature learning (HDFL) are used in [151]. Hierarchical boosting network learned features are used in [152]. The Gabor-based network learned features are used in [101]. Joint feature distribution alignment learning (JFDAL) is used in [92]. Relational Graph Module (RGM) learned features are used in [22].

e) **Usage of Boosting:** Using this tool, we can combine several weak classifiers to make strong classifiers. Here strong classifier refers to a generalised classifier. The hierarchical boosting network is used in [152]. Hierarchical discriminant feature learning is used in [151].

f) **New model architecture:** In machine learning, generalisation is achieved by introducing more complex models. Mixed adversarial examples and logits replay (MAELR) [132], Relational graph module (RGM) [22], Gabor-based network [101], Cross-modality discriminator network (CMDN) [16] are proposed for cross-domain face recognition. Deep learning-based hybrid architecture is proposed in [34]. In [105], one encoder-decoder generator and two discriminators are used for the synthesis. In [106], bidirectional heterogeneous prototype

learning architecture is proposed. A disentangled spectrum variations network is proposed in [53].

Table 2.2 provides the summary of generalisation approches used for cross-domain face recognition.

2.3 Contributions in cross-domain face recognition

In our work, we have done cross-domain face recognition on thermal to visual face recognition, and we have three main methods.

- Deep transfer learning-based method
- Common subspace learning using dictionary learning
- Collaborative metric learning-based method

- **Deep transfer learning-based method:** In this method, a separate classifier is learned for each domain. Knowledge transfer from the source domain classifier enhances the target classifier accuracy. This feature-based method utilises a dense sparse, dense training technique, with the dense sparse, dense learning approach functioning as a form of regularisation.
- **Common subspace learning using dictionary learning:** In the projection-based dictionary learning method, we have separated the data into two parts - domain-specific features and identity-related features. By considering identity-related features, we can learn common subspaces. Using this common subspace, we perform cross-domain face recognition. This method is presented in Chapter 4.
- **Collaborative metric learning based method:** We have learned a metric using maximum margin matrix factorisation in this method. In this, by improving the margin, we got a generalised metric. The proposed method is a metric learning projection-based method. This method generalises the model using maximising the margin.

Approaches for generalization	Description of Related Work
Loss fuction	In [160, 162], bi-directional dual-constrained top-ranking (BDTR) loss is proposed. In [44], dual-constrained triplet loss is introduced. In [45], hypersphere manifold embedding is learned using Sphere Softmax loss. In [84], hetero center triplet loss is proposed. Domain-based angular margin loss and a maximum angular loss are introduced in [74]. In [161], Intra-modality weighted-part aggregation loss is proposed. In [59], Maximum mean discrepancy based loss is proposed. In [33], Pairwise identity preserving loss is proposed. In [47], 3D-based pose correction loss is proposed. In [125], A model is proposed with two losses: (i) Derived-margin softmax loss (ii) Reconstruction-Center (ReCent) loss. HR-anchor loss is proposed in [127]
Regularization	In [79], sparse coupled projection method using multidimensional scaling joint L2,1-norm regularization is proposed
Maximization of margin	In [74], Angular margin loss is proposed, it maximising the margin. In [125], Derived-Margin softmax loss is proposed, it maximizing the margin.
Generalized features	SIFT, HOG, and LBP features are used in [23, 51, 120]. 2D-LDA features are used in [168]. Logarithm gradient histogram(LGH) features are used in [178], Local Quality Descriptor features are used in [8], Hierarchical discriminant feature learning(HDFL) used in [151]. Hierarchical boosting network learned features are used In [152]. Gabor-based network learned features are used in [101]. Joint feature distribution alignment learning (JFDAL) is used in [92]. Relational Graph Module (RGM) learned features are used in [22]
Boosting	Hierarchical boosting network is proposed in [152] Hierarchical discriminant feature learning is proposed in [151]
New model architecture	Mixed adversarial examples and logits replay (MAELR) architecture is proposed in [132]. Relational graph module (RGM) is proposed in [22]. Gabor-based network proposed in [101]. Cross-modality discriminator network (CMDN) is proposed in [16]. Deep learning-based hybrid architecture is proposed in [34]. In [105], A new architecture of one encoder-decoder generator and two discriminators is proposed. In [45], hypersphere manifold embedding network (HSMEnet) is proposed

Table 2.2: Cross-domain face recognition literature categories based on approaches used for the generalisation

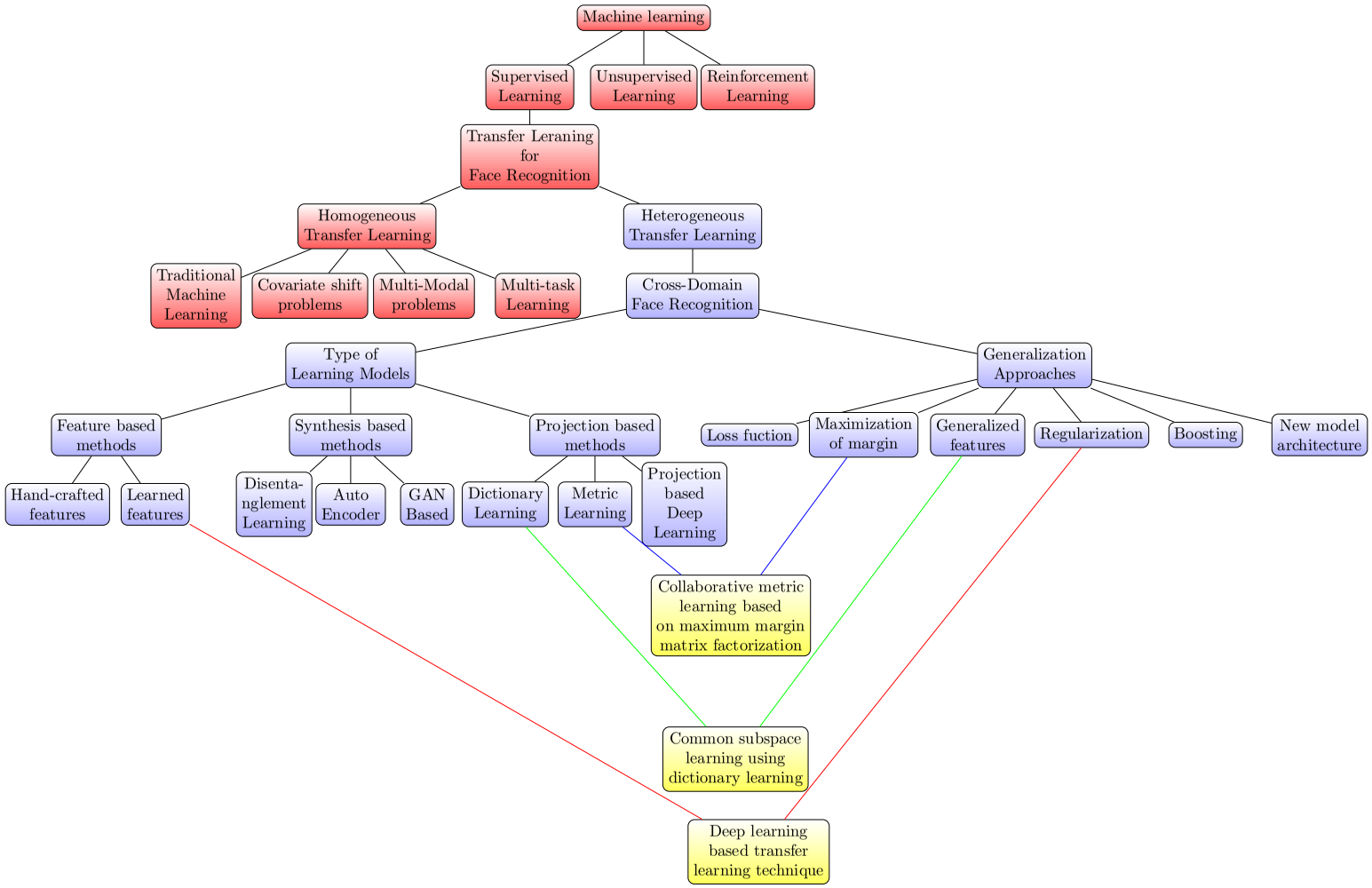


Figure 2.8: Taxonomy of literature and contribution: red nodes indicate related work, while blue nodes represent literature survey on cross-domain face recognition, and yellow nodes represent the contributions

Figure 2.8 illustrates the transfer learning for face recognition taxonomy updated with our contributions. Here, red nodes indicate related work, blue nodes represent literature survey on cross-domain face recognition, and yellow nodes represent the contributions.

2.4 Performance - Metrics

This section describes the performance metrics used for cross-domain face recognition.

- **Accuracy:** In supervised learning tasks, one way to measure how accurate a machine learning model is by its *classifier accuracy*. It is calculated by dividing the number of examples that were correctly categorised by the total number of examples in the dataset.
  Accuracy = (Number of right predictions) / (Total number of predictions)

- **Rank-k Accuracy:** Rank-k accuracy refers to the proportion of predictions made by a model that correctly identifies the true class among the top-k predicted classes.

  Rank-k accuracy = (Number of instances where the true class is in the top-k predicted classes) / (Total number of instances)

- **Equal Error Rate (EER):** Equal Error Rate (EER) is a performance metric commonly used in biometric verification systems. It is defined as the point where the false acceptance rate (FAR) equals the false rejection rate (FRR). In other words, it is the point where the rate of incorrectly accepting a false identity is the same as the rate of incorrectly rejecting a true identity.

  Incorrectly matches the input pattern to a non-matching template in the database. It measures the percentage of invalid inputs that are incorrectly accepted.

  Where FAR is the False Acceptance Rate, and FRR is the False Rejection Rate. The FAR is the probability that the system incorrectly matches the input pattern to a non-matching template in the database. It measures the percentage of invalid inputs that are incorrectly accepted, while the FRR is the probability that the system fails to detect a match between the input pattern and a matching template in the database. It measures the percentage of valid inputs that are incorrectly rejected.

  The eer is a useful metric in biometric systems because it provides a single point of comparison between two systems or algorithms without specifying a particular operating point. A lower eer indicates better performance, as it means that the system is able to balance the trade-off between the FAR and the FRR more effectively.

  The eer is typically computed using a receiver operating characteristic (ROC) curve, which plots the FAR against the FRR at different operating points. The eer can be determined by finding the point on the ROC curve where the FAR and FRR intersect.

- **Mean Average Precision (mAP):** mAP is a performance metric commonly used in information retrieval and object detection tasks to evaluate the accuracy of ranking algorithms.

  mAP is a modified version of the Average Precision (AP) metric, which is computed for each query or object and then averaged across all queries or objects in the dataset.

$$mAP = \frac{1}{N} * \sum_{i=1}^{N} AP_i$$

  Where N is the total number of queries or objects in the dataset, and AP_i is the average precision for the i^{th} query or object.

  Average Precision (AP) is a performance metric commonly used in information retrieval and object detection tasks to evaluate the accuracy of ranking algorithms.

  AP measures how well a ranking algorithm retrieves relevant items or objects by taking into account both the precision and the recall of the retrieved results.

$$AP = \sum_{k=1}^{m} \frac{(P(k) * rel(k))}{\text{(number of relevant items)}}$$

  Where m is the total number of retrieved items, $P(k)$ is the precision at rank k, $rel(k)$ is an indicator function that takes the value 1 if the k^{th} item is relevant and 0 otherwise, and the sum is overall relevant items in the dataset.

  In other words, AP computes the average of the precision values at each rank where a relevant item is retrieved, weighted by the number of relevant items.

  The mAP metric provides a more accurate measure of the overall performance of a ranking algorithm than other metrics such as accuracy or precision.

2.5 Benchmark Datasets

In this work, we have utilised three publicly available benchmark datasets: the UND-X1 dataset [31], RGB-D-T dataset [99], and RegDB dataset [98].

Table 2.3 presents the details of each benchmark dataset used. Here all three datasets' thermal images are obtained in the LWIR region.

The RGB-D-T dataset includes 51 different person images in three domains: visual, thermal, and depth. Each domain contains 15,300 (51×300) face images, and the number of images per subject is 300 for each domain. The visual image resolution is 640×480, while the thermal image resolution is 384×288.

The RegDB dataset consists of 412 people images in two different domains: visual and thermal. Each domain has 4,120 images (412 $\times$ 10). For each person, there are ten images of the visual domain and ten images of the thermal domain. It is worth noting that this dataset included images of detected individuals that were cropped and resized to 128$\times$64.

Finally, the UND-X1 dataset consists of 2,292 image pairs of 82 persons, distributed evenly in the thermal and visual domains. The number of images per subject varies from 4 to 40 for each domain. The visual image resolution is 1600$\times$1200, while the thermal image resolution is 312$\times$239.

Dataset Name	No of Subjects	No of Visual images per subject	No of Thermal images per subject	Total no of image pairs	Resolution of Visual images	Resolution of Thermal images
RegDB[98]	412	10	10	4,120	128$\times$64^{++}	128$\times$64^{++}
UND-X1[20, 31]	82	4 to 40**	4 to 40**	2,292	1,600$\times$1,200	312$\times$239
RGB-D-T[99]	51	300	300	15,300	640$\times$480	384$\times$288

Table 2.3: Details of Benchmark Datasets
$++$The RegDB dataset included images of detected individuals that were cropped and resized to 128$\times$ 64.
$**$ Number of images per subject vary from 4 to 40

2.6 Challenges of cross-domain face recognition

Following are the challenges of thermal to visual cross-domain face recognition.

1. **Domain difference:** Thermal and visual images are captured using different cameras. These cameras work on different spectrums. Thermal cameras capture the heat emitted by the objects, while visual cameras capture the reflected light from the object. This difference can result in differences in appearance, texture, and illumination between thermal and visual images, which makes it difficult to match accurately between the two domains.

2. **Availability of data:** Thermal imaging technology may not be as widely available as visual imaging technology, especially in certain environments or settings. Limited datasets are available that contain paired thermal and visual images of faces, which can be used for training and evaluating thermal to visual face recognition algorithms. This limits the availability of data for training accurate and robust algorithms, which can affect the performance of thermal to visual face recognition systems.

3. **Difference in image quality:** Thermal images typically have lower spatial resolution and lower image quality compared to visual images. Thermal cameras may

have limited resolution and sensitivity, resulting in less detailed and noisier images. As a result, important facial features may not be visible in thermal images, making it difficult to match faces with visual images accurately.

4. **Feature learning algorithms for thermal domain:** Traditional face feature learning algorithms were developed only for the visual domain. They may not be directly applicable to the thermal domain due to differences in domain characteristics. Developing a generalised and robust feature learning algorithm for the thermal domain is a challenging task.

5. **Vulnerability to adversarial attacks:** Cross-domain face recognition models can be vulnerable to adversarial attacks, where an attacker can modify an input image to cause the model to misclassify it.

In our thesis, we have addressed the first four challenges of cross-domain face recognition. Proposed methods are explicitly developed to address the domain differences between thermal and visual images and the challenges of limited data availability and differences in image quality. All three methods are automatic feature learning algorithms that can learn discriminative features from thermal and visual domains. Our techniques also helped to align the features between the two domains and improve recognition accuracy in cross-domain face recognition tasks.

CHAPTER 3

Thermal to visual face recognition using deep learning based transfer learning

In this chapter, we discussed our proposed deep transfer learning-based method for thermal to visual face recognition. For this, separate classifiers are used for each domain. In this method, we train the thermal classifier by utilising the visual classifier, which is done by using transfer learning. By using this method, the knowledge from the visual classifier gets transferred to learn the thermal classifier. As a result, this thermal classifier gets trained with even less amount of training data.

3.1 Introduction

Thermal to visual face recognition is more challenging because of the nonlinear spectral characteristics between the thermal and visual spectra. The availability of data in the thermal domain is less when compared to that of the visual domain. In thermal to visual face recognition, the visual domain is usually our source domain, and the thermal domain is our target domain.

Figure 3.1 illustrates the contribution of thermal to visual face recognition method using transfer learning. It takes the visual domain and thermal domain face images as

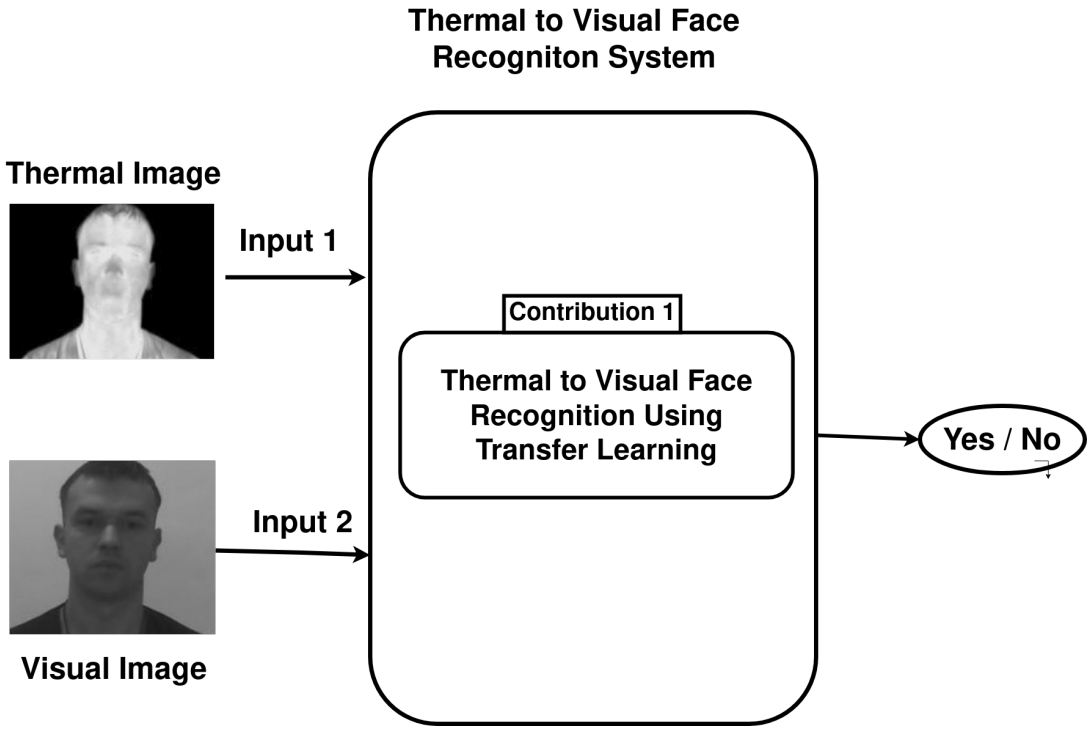


Figure 3.1: Thermal to visual cross-domain face recognition using deep learning based transfer learning technique

input and returns 'Yes' if they are of the same person's face; otherwise, it returns 'No'.

In our method, convolution neural network (CNN) [76, 77] is used for each domain, and we train a separate CNN for each domain. Let the trained CNN for the source domain be CNN_s, and the trained CNN for the target domain be CNN_t. In this, the face verification is done by two classifiers, namely CNN_s and CNN_t.

Figure 3.2(a) illustrates the visual domain classifier CNN_s. Input to the CNN_s is visual domain face image x_s, and the output is the predicted label y_s. Similarly, Figure 3.2(b) illustrates the thermal domain classifier CNN_t. Input to CNN_t is thermal domain face image x_t, and the output is predicted label y_t. Verification is done by the predicted labels in the following way. If y_s and y_t are equal, then x_s and x_t are the same person's face images, and if y_s and y_t are not equal, then x_s and x_t are different person's face images. This is illustrated in Figure 3.2(c).

Learning the thermal domain classifier (CNN_t) is the main challenge in these two classifiers methods. It is challenging because the learnability of the thermal domain image is hard, as the resolution of the thermal image is usually less when compared to that of the visual domain. To increase the learnability of the target modality classifier

by using less amount of data, we adapted the transfer learning approach [104]. Using transfer learning, we proposed two methods called DSD$_{TL1}$ and DSD$_{TL2}$. $TL1$ & $TL2$ are two mask functions. Here, DSD (Dense-Sparse-Dense) [43] is a training method that works on a single domain, whereas DSD$_{TL1}$, DSD$_{TL2}$ works for cross domain.

The rest of the chapter is organised as follows: Section 3.2 gives the related work and background. The proposed approach is given in Section 3.3. Experimental results and analysis are discussed in Section 3.4, and we summarised this contribution in Section 3.5.

3.2 Related work and Background

We use Dense-Sparse-Dense (DSD) training for our network. DSD consists of three distinct stages: 1) Initial Dense stage, 2) Sparse Dense stage, and 3) Final Dense stage. In the Initial Dense stage, the network's weights are initialised using the normal distribution, and the network undergoes training. During the Sparse Dense stage, a threshold is determined using a heuristic. This threshold serves as a criterion for pruning the weights with lower values. The criterion of the heuristic is that the number of connections pruned has to be high without losing the accuracy. The pruned network is then trained by replacing the pruned weights with zero. In the Final Dense stage, the pruned connections are recovered through re-training. The model capacity gets increased by using DSD training. In DSD$_{TL1}$ & DSD$_{TL2}$, the model capacity of the target domain classifier gets increased by leveraging the weights of the source domain classifier.

When employing transfer learning, it is important to address the following three questions [104]:

1. *When to transfer?*
   There needs to be shared knowledge between two domains. For instance, in the context of thermal to visual face recognition, a person's face image appears in both thermal and visual domains, and certain low-level features are common to both. This indicates the presence of sufficient knowledge that can be transferred between these domains.

2. *What to transfer?*
   The specific method used determines what is transferred. In our approach, one of the stages involves obtaining a sparse source network, and during this stage, the weights of this fine-tuned sparse source network are transferred to the target.

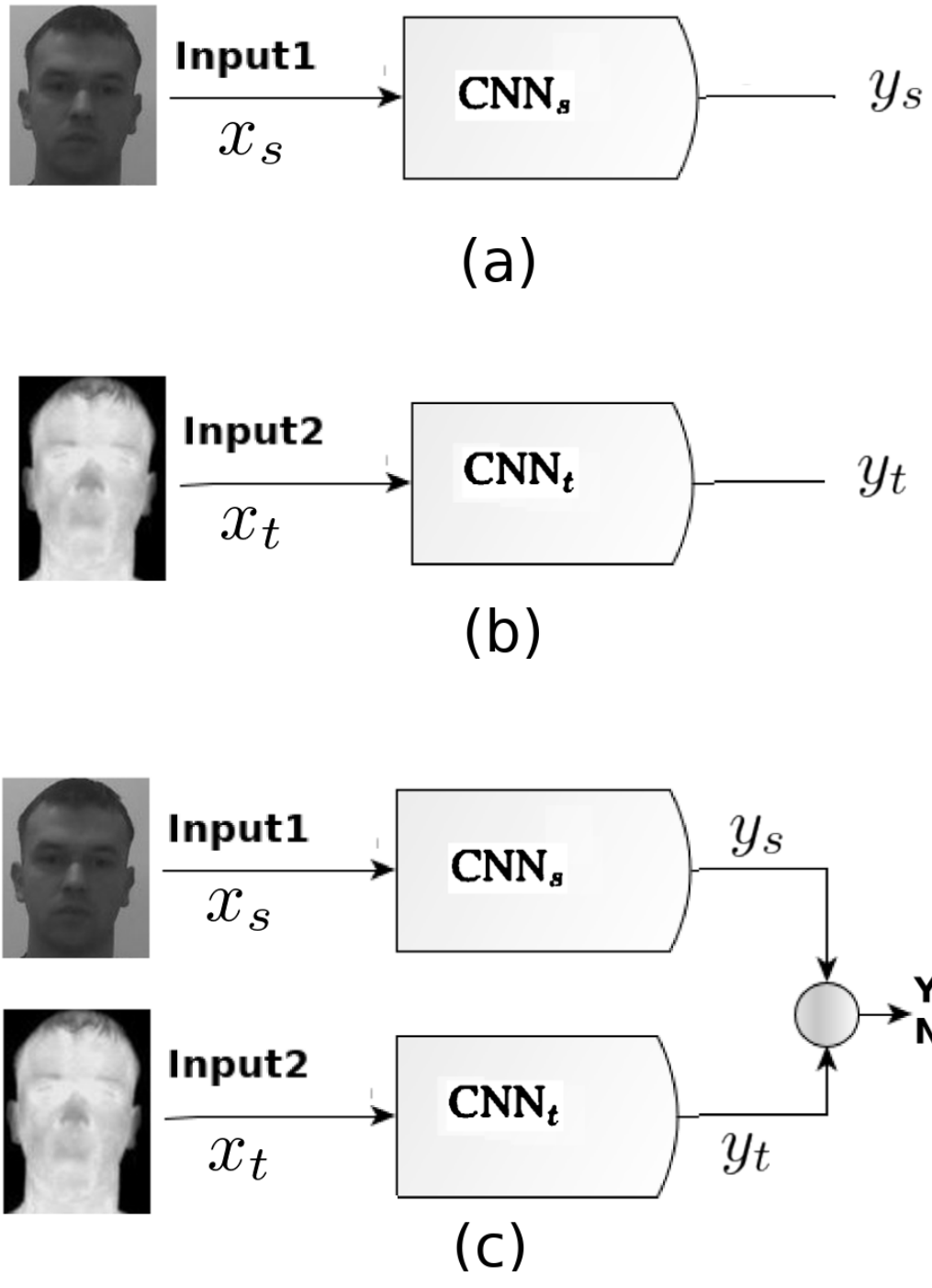


Figure 3.2: a) Visual domain classifier (CNN$_s$), y_s is predicted label b) Thermal domain classifier (CNN$_t$), y_t is predicted label c) Thermal to visual face recognition using two classifiers

3. *How to transfer?*
   The transfer process involves simply cloning the fine-tuned sparse source network as the initial sparse target network. By doing so, the knowledge and learned representations from the source network can be leveraged for the target task.

3.3 Proposed Methods: DSD$_{TL1}$ and DSD$_{TL2}$

Thermal to visual face recognition is done by two classifier methods, in which the first one is source domain classifier (CNN$_s$) and the second one is target domain classifier (CNN$_t$).

Figure 3.3 illustrates the proposed deep transfer learning method. During the training stage, Step 1 focuses on learning the visual classifier CNN$_s$, while Step 2 involves learning the thermal classifier CNN$_t$ using DSD$_{TL}$ transfer learning. In the testing stage, thermal to visual face recognition is performed using the learned classifiers.

In order to increase the accuracy of inter-modality face recognition, we need to decrease the error space (e_{intr}) of Inter-modality Face Recognition given in equation (3.1), which depends on e_s and e_t. Here, e_s is the error space of CNN$_s$ and e_t is the error space of CNN$_t$.

$$e_{intr} = e_s + e_t - (e_s \cap e_t) \tag{3.1}$$

In order to minimise e_{intr}, we need to minimise e_s and e_t and maximise $e_s \cap e_t$. In inter-modality face recognition, minimising e_t and maximising $e_s \cap e_t$ are difficult and so we are approaching transfer learning in which e_t gets minimised, and $e_s \cap e_t$ gets maximised as we are making use of source domain classifier knowledge. The method which we proposed is a transfer learning-based CNN$_t$ learning method, and we call it as DSD$_{TL}$.

Algorithm 1 illustrates Dense-Sparse-Dense Transfer Learning Method DSD$_{TL}$. $W_s^{(f)}$, X_s, X_t, λ_{TL1}, λ_{TL2}, η are inputs to the algorithm, and the output is $W_t^{(f)}$. Here,

- $W_s^{(f)}$ is the weight of the visual domain classifier CNN$_s$
- X_s is the set of source (visual) domain images $\{x_s^1, x_s^2, ..., x_s^i, ..., x_s^n\}$
- X_t is the set of target (thermal) domain images $\{x_t^1, x_t^2, ..., x_t^j, ..., x_t^m\}$

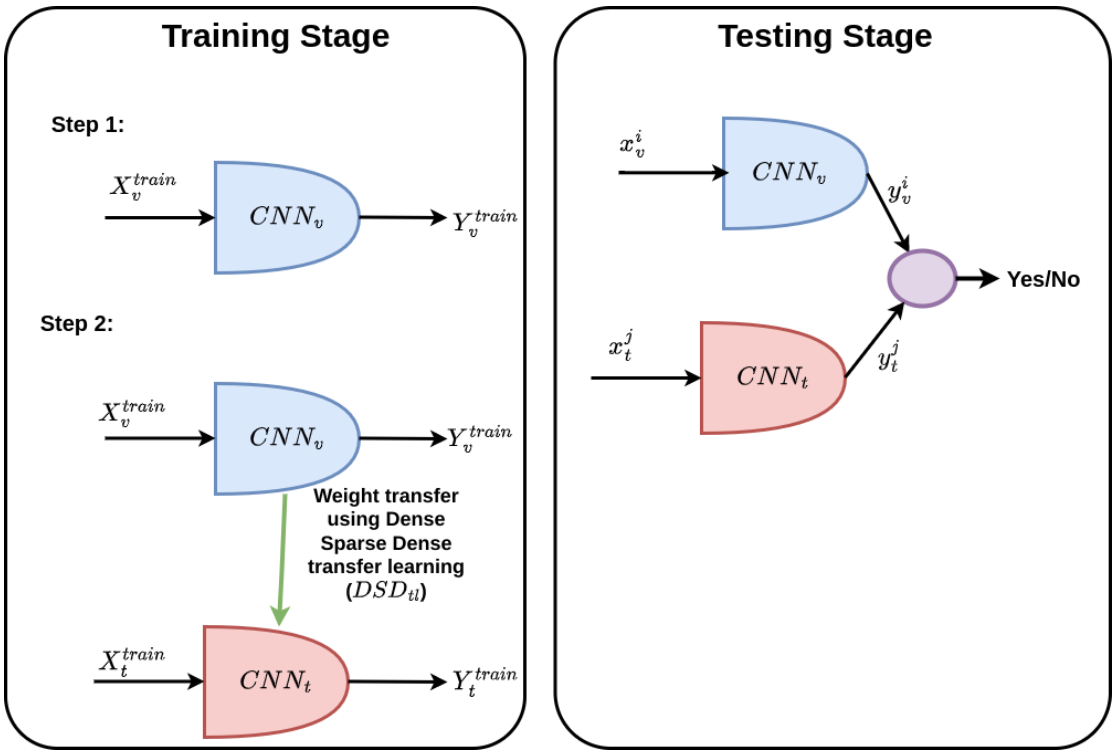


Figure 3.3: Illustration of the proposed deep transfer learning method.

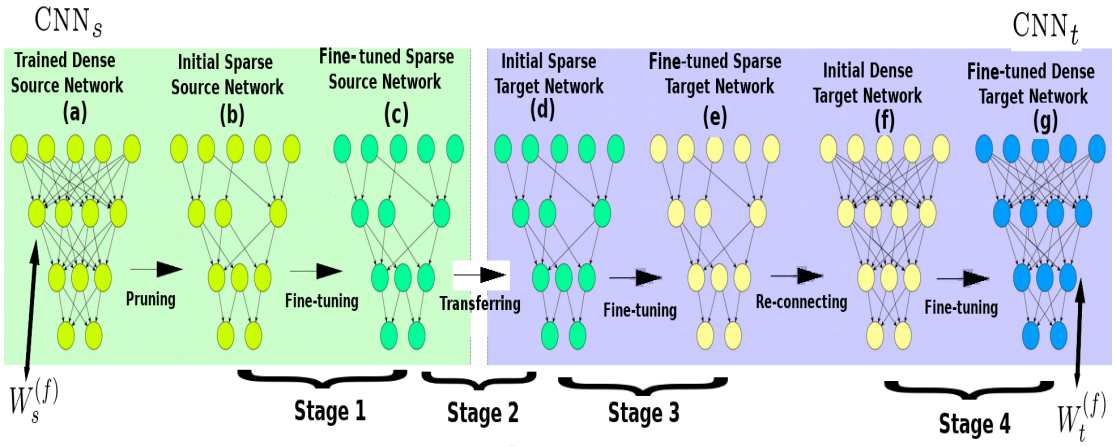


Figure 3.4: Training flow in DSD_{TL}, here (a), (b), (c) belongs to the source domain and (d), (e), (f), (g) belongs to the target domain. Weight transfer between source to target is shown at (c) to (d)

- λ_{TL1} and λ_{TL2} are threshold valued arrays
- η is an array of learning rates consisting of three values
- $W_t^{(f)}$ is the weight of the target (thermal) domain classifier CNN_t

Algorithm 2 illustrates the mask computation method TL1, which is used in Algorithm 1. A threshold (λ_{TL1}) is passed as an input. The threshold is used to prune the weights and returns a mask M1.

Algorithm 3 illustrates the mask computation method TL2, which is used in Algorithm 1. An array of two thresholds (λ_{TL2}) is passed as an input to the algorithm, and it is used to prune the weights and returns a mask M2.

Figure 3.4 illustrates the training flow of DSD$_{TL}$. The weights of CNN$_s$ exist at stage (a) of Figure 3.4. We then find the mask and prune the weights to get stage (b). By fine-tuning the network, we get stage (c). At stage (d), the weights are transferred from source to target, and we apply the same mask which we got from the source on target. By using the target data X_t, we fine-tune this network and get stage (e). Now we remove the mask and reconnect the network to get stage (f). By using X_t, we retrain the network and get stage (g). In DSD$_{TL}$, mask computation is done in two ways, as shown in Algorithm 2 and Algorithm 3. In Algorithm 2, a threshold (λ_{TL1}) is passed as an input to prune the weights, and λ_{TL2} (array of two thresholds) is passed as an input parameter to Algorithm 3 to prune the weights. Based on the method of computation of mask, there are two variations of DSD$_{TL}$ - namely DSD_{TL1} (uses mask1) and DSD_{TL2} (uses mask2). Mask1 prunes the near-zero weights only. As the near-zero valued weights are less influential for the classification of the source domain, those weights get pruned. In mask2, along with near-zero weights, the higher valued weights also get pruned, as the higher valued weights are more influential for the classification of the source domain.

3.4 Results and Analysis

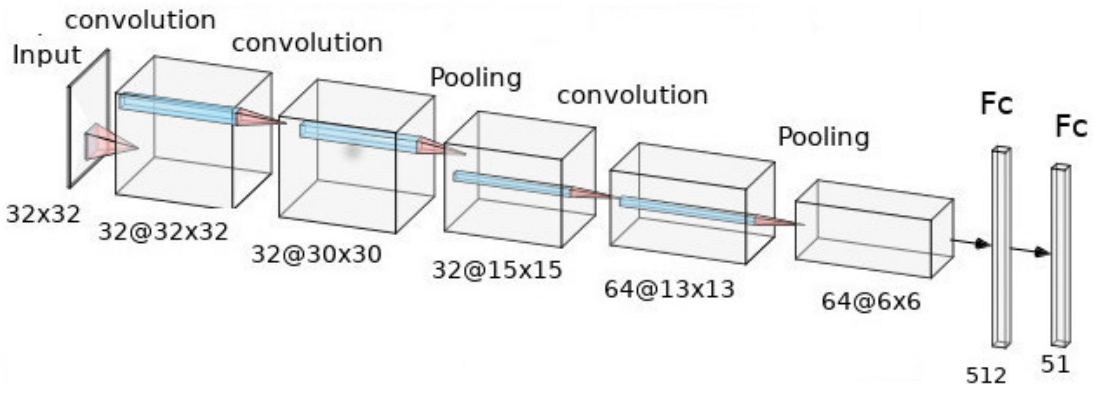


Figure 3.5: Trained CNN architecture for RGB-D-T dataset

We tested our algorithms on (i) RGB-D-T dataset [99] and (ii) UND-X1 collection [20, 31]. Both datasets, (i) and (ii), were acquired in the LWIR (Long-Wave Infrared)

Algorithm 1: Dense-Sparse-Dense Transfer Learning Method DSD$_{TLi}(i = 1, 2)$

Input : $W_s^{(f)}, X_s, X_t, \lambda_{TL1}, \lambda_{TL2}, \eta, i$

/* Where, $W_s^{(f)}$ is the set of weights of CNN$_s$, X_s is the set of source (visual) domain images, X_t is set of target (thermal) domain images , λ_{TL1} and λ_{TL2} are threshold valued arrays, η is an array of learning rates, i is mask selector. */

Output : $W_t^{(f)}$

/* Where, $W_t^{(f)}$ is the set of weights of the target (thermal) domain classifier CNN$_t$ */

if *i==1* **then**

  $M = TL1(W_s^{(f)}, \lambda_{TL1})$; // TL1 Computes the mask 1

end

if *i==2* **then**

  $M = TL2(W_s^{(f)}, \lambda_{TL2})$; // TL2 Computes the mask 2

end

$W_s^{(0)} = W_s^{(f)}$;

n=1;

/* Stage1: Pruning and Fine-tuning */

while *not converged* **do**

  $W_s^{(n)} = W_s^{(n-1)} - \eta[1]\nabla F(W_s^{(n-1)}; X_s)$; // Updation of weights

  $W_s^{(n)} = W_s^{(n)}.\, M$; // Pruning of weights

  $n = n + 1$;

end

/* Stage2: Transferring */

$W_t^{(0)} = W_s^{(f)}$;

n=1;

/* Stage3: Fine-tuning */

while *not converged* **do**

  $W_t^{(n)} = W_t^{(n-1)} - \eta[2]\nabla F(W_t^{(n-1)}; X_t)$; // Updation of weights

  $W_t^{(n)} = W_t^{(n)}.\, M$; // Pruning of weights

  $n = n + 1$;

end

n=1;

/* Stage4: Re-connecting and Fine-tuning */

while *not converged* **do**

  $W_t^{(n)} = W_t^{(n-1)} - \eta[3]\nabla F(W_t^{(n-1)}; X_t)$; // Updation of weights

  $n = n + 1$;

end

return $W_t^{(f)}$

Algorithm 2: Mask Computation Method TL1

Input : $W_s^{(f)}$, λ_{TL1}
/* Where, $W_s^{(f)}$ is the set of weights of CNN_s, λ_{TL1} is an array of threshold values. */
Output : M1
/* M1 is the mask */
/* Computation of mask M1 */
Initialization: M1 *with zeros*
(m,n)=size($W_s^{(f)}$);
for $i \leftarrow 1$ **to** m **do**
  for $j \leftarrow 1$ **to** n **do**
    if *(*$|W_s^{(f)}(i,j)| > \lambda_{TL1}(1)$*)* **then**
      M1$(i,j) = 1$; // Updation of mask
    end
  end
end
return M1

Algorithm 3: Mask Computation Method TL2

Input : $W_s^{(f)}$, λ_{TL2}
/* Where, $W_s^{(f)}$ is the set of weights of CNN_s, λ_{TL2} is an array of threshold values. */
Output : M2
/* M2 is the mask */
/* Computation of mask M2 */
Initialization: M2 *with zeros*
(m,n)=size($W_s^{(f)}$);
for $i \leftarrow 1$ **to** m **do**
  for $j \leftarrow 1$ **to** n **do**
    if *(*$\lambda_{TL2}(1) < |W_s^{(f)}(i,j)| < \lambda_{TL2}(2)$*)* **then**
      M2$(i,j) = 1$; // Updation of mask
    end
  end
end
return M2

region. The dataset details are presented in Table 3.1. The RGB-D-T dataset comprises 51 distinct person images across three domains: visual, thermal, and depth. Each domain contains 15,300 (51×300) face images, with 300 images per subject for each domain. In our experiments, we focused on thermal and visual images. Each domain is divided into two halves: one half for testing (VIS-150 test set, T-150 test set) and a portion of the other half for training (VIS-150 train set, T-150 train set).

Figure 3.5, illustrates the trained CNN architecture for the RGB-D-T dataset. The same architecture is used for the source classifier (CNN_s) and target classifier (CNN_t). The architecture has several blocks to extract features from the input data. The architecture consists of two convolutional blocks, followed by a pooling block. A convolutional block and another pooling block follow this. Finally, the architecture includes two fully connected layer blocks.

Let's briefly explain the different components of the CNN architecture mentioned:

- **Input block:-** It has the input layer of size 32 ×32. The input images are resized to 32×32 to match the network's architecture. This resizing step ensures compatibility and facilitates efficient feature extraction.
- **First convolution block:-** It contains two layers- convolution layer and activation layer.
  - **Convolution layer:**
    - Input size 32×32
    - Output size 32×32×32
    - This convolutional layer applies 32 filters of size 3×3 to the input image. Each filter detects specific low-level patterns or features in the image, and the layer's output consists of 32 feature maps, each capturing different local features. Here, zero padding ensures that the output has the same spatial dimensions as the input.
  - **Activation Layer:**
    - Input size 32×32×32
    - Output size 32×32×32
    - The Rectified linear unit (ReLU) activation function introduces nonlinearity into the network by replacing negative pixel values with zero. This helps in capturing complex patterns and features.
- **Second convolution block:-** It contains two layers- convolution layer and activation layer.
  - **Convolution layer:**
    - Input size 32×32×32
    - Output size 32×30×30

        * This convolutional layer applies 32 filters of size 3×3 to the previous layer's output. Each filter detects specific mid-level features, and the layer's output consists of 32 feature maps, each capturing different local features.

    - **Activation Layer:**

        * Input size 32×30×30
        * Output size 32×30×30
        * Similar to the previous activation layer, this layer applies the ReLU activation function to introduce non-linearity into the network.

- **First pooling block:-** It contains two layers- Max pooling layer and dropout layer.
    - **Max pooling layer**

        * Input size 32×30×30
        * Output size 32×15×15
        * The max pooling layer reduces the spatial dimensions by taking the maximum value within each 2x2 pooling region. It helps in capturing important features while reducing the computational complexity.

    - **Dropout layer:**

        * Input size 32×15×15
        * Output size 32×15×15
        * Dropout randomly sets a fraction of input units to zero during training (here, 50%). It acts as a regularisation technique to prevent overfitting by forcing the network to learn more robust features.

- **Third convolution block:-** It contains two layers- convolution layer and activation layer.
    - **Convolution layer:**

        * Input size 32×15×15
        * Output size 64×13×13
        * This convolutional layer applies 64 filters of size 3×3 to detect more complex and higher-level features from the previous layer's output.

    - **Activation Layer:**

        * Input size 64×13×13
        * Output size 64×13×13
        * ReLU activation is applied to the output of the previous convolutional layer to introduce non-linearity into the network.

- **Second pooling block:-** Contains two layers- Max pooling layer and dropout layer.
    - **Max pooling layer:**

        * Input size 64×13×13
        * Output size 64×6×6

    * Another max pooling layer with a pool size of 2x2 is applied to reduce the spatial dimensions further while retaining important features.

  – **Dropout layer:**

    * Input size 64×6×6
    * Output size 64×6×6
    * Similar to the previous dropout layer, this layer randomly sets a fraction of input units to zero (50%) during training, acting as a regularisation technique.

- **First fully connected layer block:-** It contains three layers. They are flatten layer, fully connected layer and activation layer.
  – **Flatten Layer:**

    * Input size 64×6×6
    * Output size 2,304
    * The flatten layer reshapes the 3D feature maps from the previous layer into a 1D vector. In this case, the input feature maps of size 6x6x64 are flattened into a single-dimensional vector of length 2,304. This step is necessary to connect the convolutional layers to the fully connected layers.

  – **Fully connected layer:**

    * Input size 2,304
    * Output size 512
    * This fully connected layer consists of 512 neurons. It receives the flattened vector from the previous layer as input and applies a linear transformation to produce a 512-dimensional output. This layer helps to learn high-level representations and capture global dependencies within the data.

  – **Activation layer:**

    * Input size 512
    * Output size 512
    * The ReLU activation function is applied to the previous layer's output. It introducing non-linearity to the network. It handles positive values as is, transforms negative values to zero, and enables the network to capture complex patterns and nonlinear dependencies in the data.

- **Second fully connected layer block:-** It contains three layers. They are dropout layer, fully connected layer and activation layer.
  – **Dropout layer:**

    * Input size 512
    * Output size 512
    * Dropout is applied to the previous layer output by randomly setting a fraction of input units (neurons) to zero during training (here, 50%). This regularisation technique aids in preventing overfitting and encourages the network to learn more generalised representations.

– **Fully connected layer:**

  * Input size 512
  * Output size 51
  * This is the final fully connected layer, also known as the output layer. It consists of 51 neurons, where 51 represents the number of classes or categories in the target dataset. The output of this layer is fed into an activation function to produce the final predictions.

– **Activation layer:**

  * Input size 51
  * Output size 51
  * The softmax activation function is applied to the previous layer's output. It normalises the output values into a probability distribution, assigning probabilities to each class. The class with the highest probability is considered as the predicted class by the model.

In Table 3.2, the accuracies of CNN_s (CNN for visual domain) on the VIS-150 test set are shown, in which the first column indicates the number of training images considered per subject (for example VIS-75 indicates that 75 visual images per subject are considered for training), whereas the second column gives the CNN_s accuracy. CNN_s trained using DSD is represented as $\text{CNN}_\text{s}^{\text{DSD}}$ and the accuracy of $\text{CNN}_\text{s}^{\text{DSD}}$ is given in third column.

Dataset	**Number of subjects**	**Number of visual images**	**Number of Thermal images**	**Resolution of visual images**	**Resolution of thermal images**
RGB-D-T Dataset [99]	51	15,300	15,300	640×480	384×288
UND X1 Dataset [20, 31]	82	2,292	2,292	1,600×1,200	312×239

Table 3.1: Details of datasets

Trainset	$\textbf{CNN}_\textbf{s}$	$\textbf{CNN}_\textbf{s}^{\textbf{DSD}}$
VIS-150	96.902	96.993
VIS-30	93.595	93.856
VIS-75	93.699	93.751

Table 3.2: Accuracy of source classifier on VIS-150 test set of RGB-D-T dataset

Table 3.3 presents the accuracies of CNN_t (CNN for the thermal domain) on the T-150 test set without transfer learning. The first column indicates the number of training images per subject, where, for instance, T-30 signifies that 30 thermal images per

subject were used for training. The second column displays the accuracy of CNN_t. Furthermore, we introduce $\text{CNN}_\mathbf{t}^\mathbf{DSD}$, which refers to CNN_t trained using DSD. The corresponding accuracy of $\text{CNN}_\mathbf{t}^\mathbf{DSD}$ is provided in the third column.

Trainset	$\mathbf{CNN_t}$	$\mathbf{CNN_t^{DSD}}$
T-30	82.693	82.954
T-75	89.647	89.804

Table 3.3: Accuracy of target classifier on T-150 test set of RGB-D-T dataset without transfer learning

Trainset	**Pre-trained**	**Proposed methods**	
	$\mathbf{CNN}_t^{WT}$ [102]	$\mathbf{CNN_t^{DSD_{TL1}}}$	$\mathbf{CNN_t^{DSD_{TL2}}}$
VIS-T 150-30	82.614	89.425	88.641
VIS-T 30-30	82.588	90.131	90.68
VIS-T 150-75	89.412	94.418	94.614
VIS-T 75-75	89.281	92.68	93.477

Table 3.4: Accuracy of target classifier on T-150 test set of RGB-D-T dataset using transfer learning

Table 3.4 presents the accuracies of the target CNN_t (on the T-150 test set) using transfer learning. The first column indicates the number of training images per subject from the visual and thermal domains. For example, VIS-T X-Y indicates that X images per subject are used from the visual domain to train the visual classifier, and Y images per subject are used from the thermal domain to train the thermal classifier. In the second column ($\text{CNN}t^{WT}$), the accuracy of the thermal classifier is provided when transferring weights from a pre-trained dense source network. The subsequent column displays the accuracy of the proposed methods. Under the proposed methods, one column represents the accuracy of the thermal classifier $\text{CNN}_\mathbf{t}^\mathbf{DSD_{TL1}}$ (CNN_t trained with DSD_{TL1}), and the other column corresponds to the accuracy of proposed thermal classifier $\text{CNN}_\mathbf{t}^\mathbf{DSD_{TL2}}$ (CNN_t trained with DSD_{TL2}). Here, if we consider VIS-T 150-30 in $\text{CNN}_\mathbf{t}^\mathbf{DSD_{TL1}}$ or $\text{CNN}_\mathbf{t}^\mathbf{DSD_{TL2}}$, CNN_s is trained on VIS-150 and the knowledge is transferred to CNN_t which is then trained on T-30. By examining the results from Table 3.3 and Table 3.4, it is clear that the accuracy of the target domain improves using the proposed transfer learning methods, namely DSD_{TL1} and DSD_{TL2} for different dataset sizes. When comparing DSD_{TL1} and DSD_{TL2}, DSD_{TL1} performs better when the num-

ber of visual domain images are more than thermal domain images. On the other hand, DSD_{TL2} outperforms DSD_{TL1} in all other cases.

In Table 3.5, the comparison of eer (equal error rate) is given. In this, the first column contains different methods for thermal domain classifiers. The method SVM-LBP [124] uses LBP features where the SVM gets trained on LBP features for thermal face recognition, whereas in SVM-HOG [124] SVM is trained on HOG features. In SVM-HOGOM [124] SVM gets trained on HOGOM features for thermal face recognition. If we observe the values of eer, $\text{CNN}_{\mathbf{t}}^{\mathbf{DSD_{TL1}}}$ and $\text{CNN}_{\mathbf{t}}^{\mathbf{DSD_{TL2}}}$ are performing better (lower the value better the performance).

In Table 3.6, the accuracies of thermal to visual face recognition are shown. This is tested on the T-150 test set and VIS-150 test set by considering the pairs from both sets. The number of pairs considered is 600 per subject, of which 300 pairs are positive and 300 are negative per subject. The first column of the table (VIS-T 150-30) says that we divided the dataset in such a way that 150 images per subject are considered from the visual domain (VIS), and 30 images per subject are considered from the thermal domain (T) for training. The second column gives the accuracies of inter-modality face recognition using baseline methods. The column $\text{CNN}_{\mathbf{s}}$-$\text{CNN}_{\mathbf{t}}$ under the baseline methods gives the accuracy using CNN_s on the visual domain and CNN_t on the thermal domain, whereas the next column gives the accuracies by using CNN_s and CNN_t that are optimised with DSD. The column $\text{CNN}_{\mathbf{s}}^{\mathbf{DSD}}$-$\text{CNN}_{\mathbf{t}}^{\mathbf{WT}}$ in the baseline methods gives the accuracies when CNN_s optimised with DSD is used on the visual domain, and CNN_t^{WT} which is trained by transferring the weights from dense source network is used on the thermal domain. The two columns under the proposed methods give the accuracies of inter-modality face recognition by using transfer learning (one of them is using DSD_{TL1} and the other is DSD_{TL2}) on the target domain. Consider the entry of $\text{CNN}_{\mathbf{s}}^{\mathbf{DSD}}$ - $\text{CNN}_{\mathbf{t}}^{\mathbf{DSD_{TL2}}}$ corresponding to VIS-T 150-75, which gives the accuracy of inter-modality face recognition when thermal and visual image pair is given as input, in such a way that the visual face image is given to $\text{CNN}_{\mathbf{s}}^{\mathbf{DSD}}$ and thermal face image is given to $\text{CNN}_{\mathbf{t}}^{\mathbf{DSD_{TL2}}}$.

We also experimented our methods on UND-X1 dataset [31]. This dataset consists of 4584 images of 82 subjects which are distributed evenly in the thermal and visual

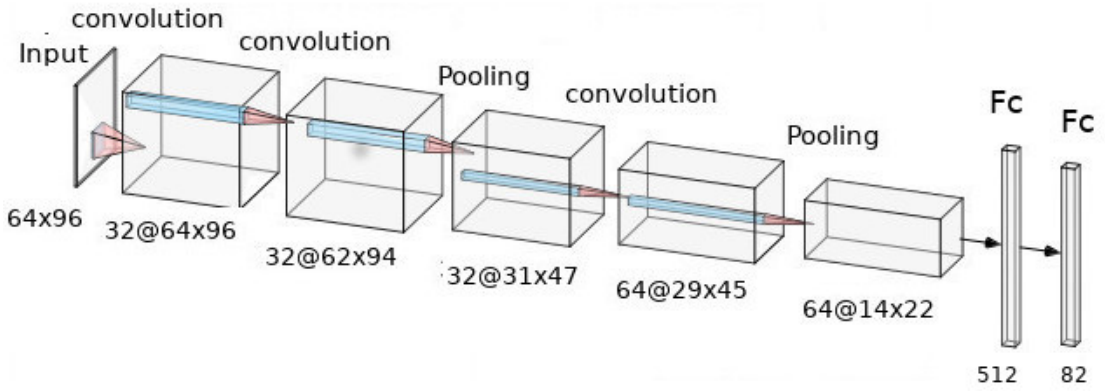


Figure 3.6: Trained CNN architecture for UND-X1 dataset

	Trainset	
Method	**T-30**	**T-75**
SVM-LBP [124]	0.36	0.133
SVM-HOG [124]	0.36	0.133
SVM-HOGOM [124]	0.36	0.066
CNN_t	0.134	0.078
$\text{CNN}_\text{t}^{\text{DSD}}$	0.132	0.075
$\text{CNN}_\text{t}^{\text{DSD}_{\text{TL1}}}$	0.072 (VIS-T 30-30)	0.0511 (VIS-T 75-75)
$\text{CNN}_\text{t}^{\text{WT}}$	0.136	0.083
$\text{CNN}_\text{t}^{\text{DSD}_{\text{TL2}}}$	**0.068** (VIS-T 30-30)	**0.043** (VIS-T 75-75)

Table 3.5: Comparison of eer on T-150 test set of RGB-D-T dataset

domains. The number of images per subject varies from 4 to 40 for each domain. The dataset is divided into a training set (70%) and a test set (30%).

Figure 3.6, illustrates the trained CNN architecture for the UND X1 dataset. The same architecture is used for the source classifier (CNN_s) and target classifier (CNN_t). The architecture has several blocks to extract features from the input data. The architecture consists of two convolutional blocks, followed by a pooling block. A convolutional block and another pooling block follow this. Finally, the architecture includes two fully connected layer blocks.

Let's briefly explain the different components of the CNN architecture mentioned:

- **Input block:-** It has the input layer of size 64 $\times$96. The input images are resized to 64$\times$96 to match the network's architecture. This resizing step ensures compatibility and facilitates efficient feature extraction.
- **First convolution block:-** It contains two layers- convolution layer and activation layer.

Trainset	Baseline methods			Proposed methods	
	$\mathbf{CNN_s}$-$\mathbf{CNN_t}$	$\mathbf{CNN_s^{DSD}}$-$\mathbf{CNN_t^{DSD}}$	$\mathbf{CNN_s^{DSD}}$-$\mathbf{CNN_t^{WT}}$	$\mathbf{CNN_s^{DSD}}$-$\mathbf{CNN_t^{DSD_{TL1}}}$	$\mathbf{CNN_s^{DSD}}$-$\mathbf{CNN_t^{DSD_{TL2}}}$
VIS-T 150-30	82.458	82.693	82.484	**89.15**	88.366
VIS-T 30-30	88.288	82.627	82.471	89.673	**90.248**
VIS-T 75-75	89.255	89.386	89.02	92.301	**93.085**
VIS-T 150-75	89.307	89.464	89.15	94.105	**94.327**

Table 3.6: Accuracy of thermal to visual face recognition on RGB-D-T dataset

$\mathbf{CNN_s}$	$\mathbf{CNN_s^{DSD}}$	$\mathbf{CNN_t}$	$\mathbf{CNN_t^{DSD}}$
94.186	95.058	83.14	84.302

Table 3.7: Accuracy of target domain classifier on UND X1 dataset without transfer learning

  - **Convolution layer:**
    - Input size 64×96
    - Output size 64×96×
    - This convolutional layer applies 32 filters of size 3×3 to the input image. Each filter detects specific low-level patterns or features in the image, and the layer's output consists of 32 feature maps, each capturing different local features. Here, zero padding ensures that the output has the same spatial dimensions as the input.
  - **Activation Layer:**
    - Input size 32×64×96
    - Output size 32×64×96
    - The Rectified linear unit(ReLU) activation function introduces non-linearity into the network by replacing negative pixel values with zero. This helps in capturing complex patterns and features.

- **Second convolution block:-** It contains two layers- convolution layer and activation layer.
  - **Convolution layer:**
    - Input size 32×64×96
    - Output size 32×62×94
    - This convolutional layer applies 32 filters of size 3×3 to the previous layer's output. Each filter detects specific mid-level features, and the layer's output consists of 32 feature maps, each capturing different local features.
  - **Activation Layer:**
    - Input size 32×62×94
    - Output size 32×62×94
    - Similar to the previous activation layer, this layer applies the ReLU activation function to introduce non-linearity into the network.

- **First pooling block:-** It contains two layers- Max pooling layer and dropout layers.

Pre-trained	Proposed methods	
$\mathbf{CNN_t^{WT}}$ [102]	$\mathbf{CNN_t^{DSD_{TL1}}}$	$\mathbf{CNN_t^{DSD_{TL2}}}$
83.721	86.047	91.86

Table 3.8: Accuracy of target domain classifier on UND X1 dataset with transfer learning

	Baseline methods			Proposed methods	
	$\mathbf{CNN_s}$-$\mathbf{CNN_t}$	$\mathbf{CNN_s^{DSD}}$-$\mathbf{CNN_t^{DSD}}$	$\mathbf{CNN_s^{DSD}}$-$\mathbf{CNN_t^{WT}}$	$\mathbf{CNN_s^{DSD}}$-$\mathbf{CNN_t^{DSD_{TL1}}}$	$\mathbf{CNN_s^{DSD}}$-$\mathbf{CNN_t^{DSD_{TL2}}}$
UND X1 Dataset	81.54	82.485	82.122	83.94	**90.334**

Table 3.9: Accuracy of thermal to visual face recognition on UND X1 dataset

  - **Max pooling layer:**
    - Input size 32×62×94
    - Output size 32×31×47
    - The max pooling layer reduces the spatial dimensions by taking the maximum value within each 2x2 pooling region. It helps in capturing important features while reducing computational complexity.
  - **Dropout layers:**
    - Input size 32×31×47
    - Output size 32×31×47
    - Dropout randomly sets a fraction of input units to zero during training (here, 50%). It acts as a regularisation technique to prevent overfitting by forcing the network to learn more robust features.

- **Third convolution block:-** It contains two layers- convolution and activation layers.
  - **Convolution layer:**
    - Input size 32×31×47
    - Output size 64×29×45
    - This convolutional layer applies 64 filters of size 3×3 to detect more complex and higher-level features from the previous layer's output.
  - **Activation Layer:**
    - Input size 64×29×45
    - Output size 64×29×45
    - ReLU activation is applied to the output of the previous convolutional layer to introduce non-linearity into the network.

- **Second pooling block:-** contains two layers: Max pooling layer and the dropout layer.
  - **Max pooling layer:**
    - Input size 64×29×45
    - Output size 64×14×22

* Another max pooling layer with a pool size of 2x2 is applied to reduce the spatial dimensions further while retaining important features.

– **Dropout layers:**

* Input size $64 \times 14 \times 22$
* Output size $64 \times 14 \times 22$
* Similar to the previous dropout layer, this layer randomly sets a fraction of input units to zero (50%) during training, acting as a regularisation technique.

- **First fully connecte layer block:-** It contains three layers. They are flatten layer, fully connected layer and activation layer.
  – **Flatten Layer:**
    * Input size $64 \times 14 \times 22$
    * Output size 19,712
    * The flatten layer reshapes the 3D feature maps from the previous layer into a 1D vector. In this case, the input feature maps of size $64 \times 14 \times 22$ are flattened into a single-dimensional vector of length 19,712. This step is necessary to connect the convolutional layers to the fully connected layers.
  – **Fully connected layer:**
    * Input size 19,712
    * Output size 512
    * This fully connected layer consists of 512 neurons. It receives the flattened vector from the previous layer as input and applies a linear transformation to produce a 512-dimensional output. This layer helps to learn high-level representations and capture global dependencies within the data.
  – **Activation layer:**
    * Input size 512
    * Output size 512
    * The ReLU activation function is applied to the previous layer's output. It introduces non-linearity to the network. It does not change the positive values and transforms negative values to zero, and enables the network to capture complex patterns and nonlinear dependencies in the data.

- **Second fully connected layer block:-** It contains three layers. They are dropout layer, fully connected layer and activation layer.
  – **Dropout layer:**
    * Input size 512
    * Output size 512
    * Dropout is applied to the previous layer output by randomly setting a fraction of input units (neurons) to zero during training (here, 50%). This regularisation technique aids in preventing overfitting and encourages the network to learn more generalised representations.

- **Fully connected layer:**
  - Input size 512
  - Output size 82
  - This is the final fully connected layer, also known as the output layer. It consists of 82 neurons, where 82 represents the number of classes or categories in the target dataset. The output of this layer is fed into an activation function to produce the final predictions.
- **Activation layer:**
  - Input size 82
  - Output size 82
  - The softmax activation function is applied to the previous layer's output. It normalises the output values into a probability distribution, assigning probabilities to each class. The class with the highest probability is considered as the predicted class by the model.

In Table 3.7, the accuracies of target domain classifiers on UND X1 dataset without transfer learning are given. CNN_s and CNN_s^{DSD} are source domain classifiers and are tested on the UND-VIS test set, and the corresponding accuracies are given in the first two columns of Table 3.7. CNN_t and CNN_t^{DSD} are target classifiers, and these are tested on the UND-T test set, and the accuracies are given in the third and fourth columns of Table 3.7.

In Table 3.8, the accuracies of target domain classifiers on UND X1 dataset with transfer learning are given. in which the first column (CNN_t^{WT}) gives the accuracy of the thermal classifier by transferring the weights from pre-trained dense source network. The next column gives the accuracy of proposed methods ($\text{CNN}_{\mathbf{t}}^{\mathbf{DSD_{TL1}}}$ and $\text{CNN}_{\mathbf{t}}^{\mathbf{DSD_{TL2}}}$) which are having better accuracies on target domain.

Table 3.9 presents the accuracies of thermal to visual face recognition on the UND X1 dataset. The evaluation is conducted on the UND-VIS test set and the UND-T test set, considering pairs from both sets. The total number of pairs considered for testing is 2,752, with 1,376 positive pairs and 1,376 negative pairs. In the second column, the accuracies of inter-modality face recognition using baseline methods are provided. The column labelled CNNs-CNN$\mathbf{t}$ indicates the accuracies obtained by using CNN_s for the visual domain and CNN_t for the thermal domain. The subsequent column displays the accuracies achieved by optimising CNN_s and CNN_t with the Dense-Sparse-Dense (DSD) training approach. The two columns under the proposed methods present the accuracies of inter-modality face recognition obtained by applying transfer learning

techniques. One of the proposed methods utilises DSD_{TL1} for transfer learning, while the other employs DSD_{TL2} for transfer learning specifically in the target domain.

In our experiments we have used Keras library [24] on top of TensorFlow library [2]. In both the architectures (Figure 3.5 & 3.6), we have used the ReLU activation function in the first four layers and the softmax function in the final layer. The optimisation used is RMSProp, and the loss used is categorical cross-entropy. The learning rate range (η) varies from 0.001 to 0.1. The λ_{TL1} range from 0.05 to 1.5. The range of $\lambda_{TL2}(1)$ ranges from 0.05 to 1.5 and that of $\lambda_{TL2}(2)$ varies from 2.5 to 4.5. The number of training epochs ranges from 1,000 to 5,000. The convergence criterion can be either the number of epochs or no decrease in loss, whichever is earlier. All the experiments were carried out on the NVIDIA Tesla M40 system. A single epoch on the GPU system takes 30-50 seconds.

3.5 Summary

We have presented a thermal to visual face recognition approach utilising a two-classifier method. Initially, the source classifier is trained, and subsequently, leveraging the knowledge of the source classifier, the target classifier is learned. The model capacity of target classifier (CNN_t) is enhanced by using the proposed methods ($\text{CNN}_{\mathbf{t}}^{\mathbf{DSD_{TL1}}}$ and $\text{CNN}_{\mathbf{t}}^{\mathbf{DSD_{TL2}}}$), in which the fine-tuned weights of sparsified source network gets transferred.

The accuracies obtained from both methods demonstrate their robust performance in the thermal to visual face recognition task. This method is tested on RGB-D-T dataset (45900 images) and UND-X1 collection (4584 images). Experimental results show that the overall accuracy of thermal to visual face recognition by transferring the knowledge gets increased from 89.3% to 94.32% on RGB-D-T dataset and from 81.54% to 90.33% on UND-X1 dataset.

CHAPTER 4

Common subspace learning for thermal to visual face recognition using dictionary learning

In the previous chapter, we discussed thermal to visual face recognition using transfer learning, and the method works only for closed set problems. In the closed-set problems, training and testing will be done on the same person's images. However, in the practical scenario, we may have to match the unknown person's face images. To address this challenge, we have proposed thermal to visual face recognition using dictionary learning. In this chapter, we have discussed the proposed dictionary learning-based thermal to visual face recognition.

4.1 Introduction

In this chapter, we have proposed thermal to visual face recognition, which has two stages. In the first stage, we project both domain images into a common subspace. This common subspace is learned using dictionary learning in which each of the face images is represented with the representation code. In the second stage, metric learning is done to measure the similarity between corresponding common subspace representations.

Figure 4.1 illustrates the contribution of common subspace learning thermal to visual cross domain face recognition method using dictionary learning. It takes the visual

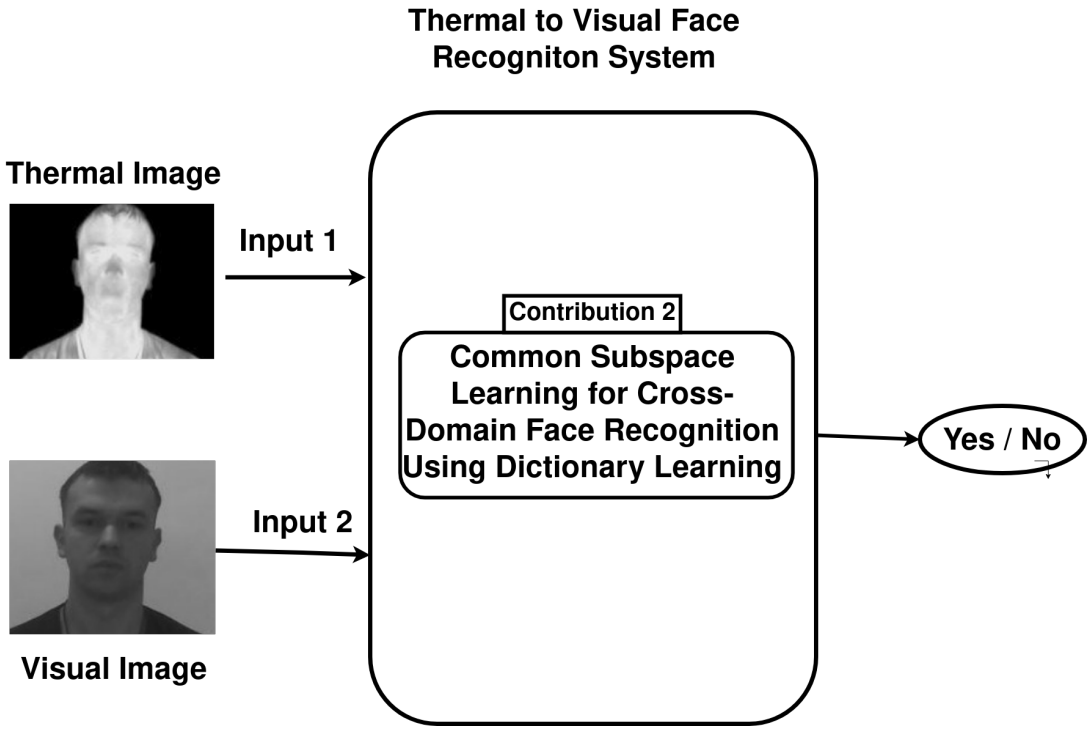


Figure 4.1: Thermal to visual cross domain face recognition based on common subspace learning using dictionary learning

domain and thermal domain face images as input and returns 'Yes' if they are of the same person's face; otherwise, it returns 'No'.

4.1.1 Dictionary Learning

Dictionary Learning is a representation learning technique. This dictionary learning is extensively used in signal-processing applications. Applications such as compressive sensing, signal de-noising, image super-resolution, and signal classification have become increasingly popular. The input signal data can be written in linear combinations of basis vectors (atoms), and the set of those basis vectors is called a dictionary. The goal of dictionary learning is to minimize the following:

$$\min_{\mathcal{D},\alpha} ||\mathbf{X} - \mathcal{D}\alpha||_2^2 + \lambda||\alpha||_1$$

where λ is the regularisation parameter which is used to balance the reconstruction error and the amount of sparsity induced by the l_1 penalty. Since most signals are conveyed by a linear combination of just a few number of basis vectors, dictionary learning is often referred to as representation code. The number of atoms and the sparsity level is

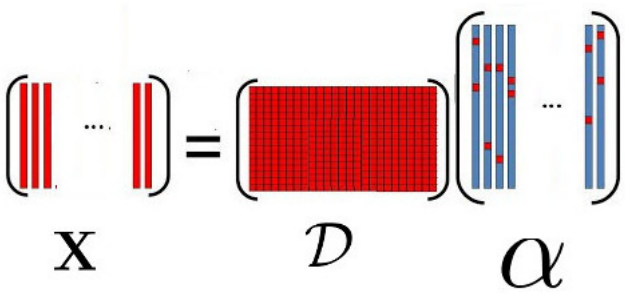


Figure 4.2: $\mathbf{X}_{m\times n}$ is the set of data vectors each of dimension m, $\mathcal{D}_{m\times k}$ is the set of basis vectors, and $\alpha_{k\times n}$ is the set of representation coefficients

important in dictionary learning [112].

Figure 4.2 illustrates the dictionary and representation code. Here $\mathbf{X}_{m\times n}$ is a set of data vectors each of dimension m, $\mathcal{D}_{m\times k}$ is a set of basis vectors each of dimension k, and $\alpha_{k\times n}$ is set of representation coefficients.

4.1.2 Dictionary Learning for Face Recognition

Dictionary learning is used for face recognition. There are five categories of dictionary learning for face recognition [153]. Figure 4.3 illustrates the five categories of dictionary learning for face recognition, which are as follows: shared dictionary learning, class-specific dictionary learning, auxiliary dictionary training, commonality and particularity dictionary learning, and domain adaptive dictionary learning.

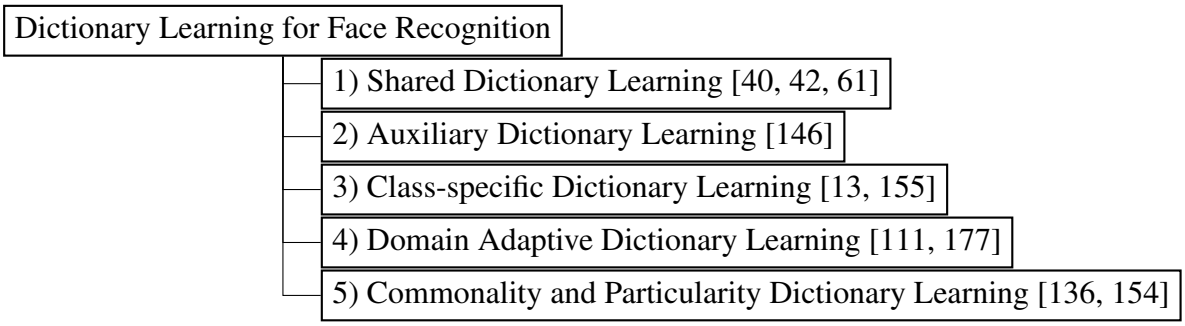


Figure 4.3: Categories of dictionary learning for face recognition

1. **Shared Dictionary Learning:** In shared dictionary learning, a common dictionary is learned for all classes of the training set. In this, classification is done by learned representation coefficients. Shared dictionary learning is further divided into two categories, namely - Label constrained model and Locality constrained model.
   - **Label constrained model:** The main aim of the label-constrained model is to improve the discriminative ability of the dictionary by using labels of

training. LC-KSVD [61] improves the discriminative ability of the shared dictionary. The objective function of the LC-KSVD algorithm is as follows.

$$\min_{\mathcal{D},\alpha,W,A} ||\mathbf{X} - \mathcal{D}\alpha||_2^2 + \beta_1||H - W\alpha||_2^2 + \beta_2||Q - A\alpha||_2^2 + \lambda||\alpha||_1$$

where $||H - W\alpha||_2^2$ is classification error term. W is the classifier parameter. $H_{c\times n}$ is label matrix of training samples. $||Q - A\alpha||_2^2$ is the discriminative sparse-code error term. Q is discriminative sparse code corresponding to training data. A is the linear transformation matrix.

- **Locality constrained model:** Each data point is represented using a sparse linear combination of dictionary atoms. The objective is to ensure that the locality of its k-nearest neighbours with the same label is preserved.

$$\min_{\mathcal{D},\alpha} ||\mathbf{X} - \mathcal{D}\alpha||_2^2 + \beta_1 \sum_{i\in\mathcal{N}^+(j)} ||\alpha_i - \alpha_j||_2^2 - \beta_2 \sum_{i\in\mathcal{N}^-(j)} ||\alpha_i - \alpha_j||_2^2 + \lambda||\alpha||_1$$

where $\mathcal{N}^+(j)$ denotes the k-nearest neighbours of α_j and same class of α_j. $\mathcal{N}^-(j)$ denotes the k-nearest neighbours of α_j and other class of α_j.
Haghiri [42] presented a discriminative dictionary learning algorithm that preserved the local structure of the training samples. Kernel collaborative representation classification with locality-constrained dictionary (KCRC-LCD) is proposed by Liu [85], in which the locality information of training samples and atoms is preserved.

2. **Auxiliary dictionary learning:** This approach improves the classification performance when each class has limited training samples.

3. **Class-specific dictionary learning** The class-specific dictionary learning algorithm is usually designed to capture the main characteristics of each class. In class-specific dictionary learning, each class has its own dictionary.

$$\min_{\mathcal{D},\alpha} \sum_{i=1}^{C} \left(||X - \mathcal{D}\alpha||_2^2 + ||X^i - \mathcal{D}^i\alpha||_2^2 \right) + \beta_1||\alpha||_1$$

4. **Domain adaptive dictionary learning:** In domain adaptive dictionary learning, a transformation matrix is estimated along with the dictionary to minimise the domain discrepancy. The objective function is defined as,

$$\begin{aligned}\min_{\mathcal{D}_t,\mathcal{D}_s,\alpha_t,A,W} &||X_t - \mathcal{D}_t\alpha_t||_2^2 + ||X_sA^T - \mathcal{D}_s\alpha_t||_2^2 \\ &+\beta_1||Q - B\alpha_t||_2^2 + \beta_2||H - W\alpha_t||_2^2\end{aligned}$$

Where Q is discriminative sparse code corresponding to training data. A is the linear transformation matrix, α_t is training sample of target domain. Coupled dictionary learning method [114] is a domain adaptation dictionary learning method. The objective function is,

$$\min_{\mathcal{D}_t,\mathcal{D}_s,\alpha_t,A,W} ||X_t - \mathcal{D}_t\alpha_t||_2^2 + ||X_s - \mathcal{D}_s\alpha_s||_2^2 - ||\alpha_s - \alpha_t||_2^2$$

5. **Commonality and particularity dictionary learning:** There are two dictionaries in commonality and particularity dictionary learning - a particularity dictio-

nary and a commonality dictionary. The particularity dictionary contains class-specific features, whereas the commonality dictionary contains features that are common to all classes. In this method, classification is done by removing the commonality. In the literature [70] [134], there are few attempts to separate the commonality and particularity of objects. This method is an extension of class-specific and shared dictionary learning methods. Like in class-specific dictionary learning, it learns separate dictionary ($\mathcal{D}_i$, i=1...C) for each class. Like a shared dictionary, it has a common shared dictionary ($\mathcal{D}_0$). By optimising the following objective function, commonality features go into $\mathcal{D}_0$ and particularity features go into $\mathcal{D}_i$.

$$f(\mathcal{D}, \alpha, \mathcal{D}_0, \alpha^0) = ||X - \mathcal{D}_0\alpha^0 - \mathcal{D}\alpha||_F^2 + \sum_{i=1}^{C}(||X_i - \mathcal{D}_0\alpha_i^0 - \mathcal{D}_i\alpha_i^i||_F^2 + \sum_{j\neq i} ||\mathcal{D}_j\alpha_i^j||_F^2) + \lambda||\alpha||_1 + \lambda||\alpha^0||_1$$

In the proposed method, common subspace learning is done using commonality and particularity dictionary learning. Using this dictionary learning, we have separated domain-specific features.

4.1.3 Metric learning

Metric learning is a subset of machine learning in which the function is learned to estimate the similarity/distance of two objects. From the input data, a function will be learned. In comparison to standard distance measurements [73], metric learning approaches are more resilient because similarity is learned from data. On the other hand, we also used deep metric learning. Deep learning-based metric learning is one of the major categories in metric learning. Deep metric learning uses deep architectures to obtain the latent space features' similarity through non-linear subspace learning. It is learned in mainly two ways - Siamese [50] network and Triplet [48] network. The Siamese network learns pair-wise similarity by using positive pairs and negative pairs. Triplet network architecture learns the similarity using positive and negative pairs simultaneously. It creates triplets comprising an anchor image, a positive image (identical to the anchor image), and a negative image (which is dissimilar to the anchor image). Deep metric learning comprises two main stages: feature learning and metric learning. In this work, feature learning is done using convolution layers, and metric learning is done using contrastive loss [41].

In this chapter, we present a two-stage approach for thermal to visual cross-domain

face recognition, based on dictionary learning. In the first stage, we aim to project images from both domains onto a common subspace. To achieve this, we represent the face images using representation codes and corresponding dictionary atoms. In the second stage, we focus on metric learning, which measures the similarity between corresponding representations in the common subspace. To find the common subspace, we employ commonality and particularity dictionary learning. In our approach, we treat all visual face images as one class and all thermal face images as another class. By leveraging commonality and particularity dictionary learning, we can separate the common features shared between the classes (visual and thermal) and the specific features unique to each class. In this context, it's important to note that the same person's face image is present in both classes (domains), and the common features between the two classes are related to the person's identity. These identity-related features are captured in the commonality dictionary, while the domain-specific features are captured in the particularity dictionary. This approach differs from conventional single-domain commonality and particularity dictionary learning, where identity-related features are typically found in the particularity dictionary. By eliminating the domain-specific features from the input data, we obtain a common subspace that preserves the person's identity. After projecting the images into this common subspace, we learn the representation using the commonality dictionary and its corresponding representation codes. This learned representation is used to find the similarity between two images, where each image is represented with one representation code. To estimate the similarity between representation codes in our proposed approach, we employ two metric learning methods. The first method is large-scale metric learning (LSML) [73], and the second metric learning approach we utilise is deep metric learning, specifically using a siamese network [50].

4.2 Related work and Background

Early research methods in cross-domain face recognition often relied on handcrafted features, which falls under the category of feature-based methods. Shuowen *et al.*[51] proposed a method that extracts handcrafted features like SIFT, HOG, and LBP for both visual and thermal domain images. These extracted features were then matched using partial least square discriminant analysis (PLS-DA). Sarfraz *et al.*[120] proposed a method where SIFT features from thermal and visual domain images were extracted

and matched using deep neural networks. The other category of feature-based methods is automatic feature learning methods. There are mainly two approaches in automatic feature learning methods: deep learning-based and dictionary learning-based. In terms of deep learning-based automatic feature learning, Riggan *et al.*[117] proposed a neural network-based coupled architecture network that forces the hidden layers of two neural networks to be as similar as possible. Christopher *et al.*[113] proposed a variant of contrastive loss to train a convolutional neural network-based coupled network. For dictionary learning-based automatic feature learning, there have been studies such as [55, 114] that focused on coupled dictionary learning algorithms. These algorithms aimed to learn a shared feature space for cross-domain image data.

Our work considers the commonality and particularity dictionary learning [134, 135], which is commonly used in single-domain face recognition algorithms. These algorithms typically involve learning separate dictionaries for each class, similar to class-specific dictionary learning. However, in addition to the class-specific dictionaries, a shared dictionary is also learned. The shared dictionary captures the common features that are present across all classes. These common features are utilised only for representation purposes. By separating the common features from the class-specific features, we enhance the discrimination capability of the class-specific features.

Mudunuri *et al.*[93] proposed a dictionary-aligned low-resolution and heterogeneous face recognition method. This method first learns the orthogonal dictionary for two domains and aligns the dictionary atoms based on atom correlation. With these aligned dictionaries, the method computes the aligned representation code. To find the final similarity, the authors have used metric learning on the representation code. In our method, we get the aligned representation code by projecting the data into the common subspace.

Ye *et al.*[159] proposed a dual stream network-based thermal to visual face recognition model. In this, one stream is for visual images, and the other stream is for thermal images, and both streams are learned using two losses, one of which is identity loss, and the other one is contrastive loss. These are further optimised with hierarchical cross-modality metric learning [159]. In another work, dual stream network [162] is learned using identity loss and dual-constrained top ranking. In our deep metric learning method, we employ similar methods like a two-stream network, and the learned

features have the identity information. In our method, identity features are coming from the common subspace representation code.

4.3 Proposed Methods: CSL1+LSML, CSL1+DML, CSL2+LSML, and CSL2+DML

This section presents the proposed thermal to visual face recognition methods: CSL1+LSML, CSL1+DML, CSL2+LSML, and CSL2+DML. For simplicity of notation, we use Method-1 for CSL1+LSML, Method-2 for CSL1+DML, Method-3 for CSL2+LSML, and Method-4 for CSL2+DML. These methods are designed as two-stage approaches.

- **Stage 1:** Common Subspace Learning
- **Stage 2:** Metric Learning

The proposed methods incorporate two variants of common subspace learning and two variants of metric learning. The common subspace learning variants are CSL1 and CSL2, while the metric learning variants are LSML and DML. Each of the four proposed methods follows a two-stage approach:

- **Stage 1:** Common Subspace Learning
  - CSL1
  - CSL2
- **Stage 2:** Metric Learning
  - LSML
  - DML

These proposed methods consist of four building blocks, combining the two common subspace learning methods and the two metric learning methods. The subsequent subsection explains these building blocks and comprehensively explains the proposed methods in each subsection.

4.3.1 Building Blocks of the Proposed Method

All four proposed methods comprise two stages: common subspace learning and metric learning.

4.3.1.1 Common subspace learning

Common subspace learning aims to find a new subspace in such a way that it removes the domain-specific features and preserves the identity-related features. We remove the domain-specific features because they are not helpful for recognition.

Notations:

- $X_v \in \Re^{d \times n_v}$ is set of visual domain images
- n_v is the number of visual domain images
- d is the dimensionality of the image
- $X_t \in \Re^{d \times n_t}$ be the set of thermal domain images
- n_t is the number of thermal domain images
- $X \in \Re^{d \times N}$ represent both visual and thermal image data, $X = [X_v \; X_t]$
- N represents the total number of visual and thermal images i.e., $N = n_v + n_t$
- $\hat{D} \in \Re^{d \times \hat{k}}$ is the overall dictionary having two parts D and D_0. Column wise concatenation of D and D_0 is $\hat{D}$ i.e., $\hat{D} = [D \; D_0]$
- $\hat{k} = k_v + k_t + k_0$ is the dimensionality of latent space for whole data
- k_0 is the dimensionality of latent space for common data
- k_t is the dimensionality of latent space of thermal data
- k_v is the dimensionality of latent space of visual data
- D_0 is the shared dictionary between two domains
- $D_t \in \Re^{d \times k_t}$ is the thermal domain-specific dictionary
- $D_v \in \Re^{d \times k_v}$ is the visual domain-specific dictionary
- $D \in \Re^{d \times k}$ is the domain-specific combined dictionary and the column concatenation of D_v and D_t is D, i.e., $D = [D_v \; D_t]$
- $k = k_v + k_t$ is the dimensionality of latent space for class-specific data
- $\alpha_t \in \Re^{k_t \times N}$ represents the thermal domain representation code and is learned with the thermal class-specific dictionary D_t

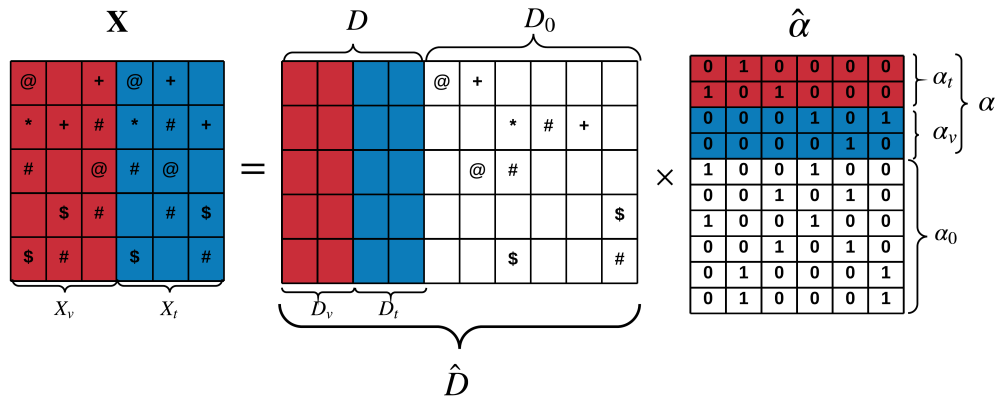


Figure 4.4: Illustration of common subspace learning. Here input data (X) is factorized into dictionary ($\hat{D}$) and representation code ($\hat{\alpha}$). Colour represents domain specific features and {@, *, #, $, +} represent identity related features. In $\hat{D}$ separated the domain related atoms and identity-related atoms

- $\alpha_v \in \Re^{k_v \times N}$ represents visual domain representation code, which will be learned with visual domain dictionary D_v
- The domain-specific representation code is $\alpha \in \Re^{k \times N}$ and $\alpha = [\alpha_v \ \alpha_t]^T$
- $\alpha_0 \in \Re^{k_0 \times N}$ represent commonality representation code, which will be learned with shared dictionary D_0
- Combined representation code is $\hat{\alpha} = [\alpha \ \alpha_0]^T$
- m, m_0, m_v, m_t be the mean vectors of $\alpha, \alpha_0, \alpha_v, \alpha_t$ respectively

Figure 4.4 illustrates the common subspace learning method. Here input data X has two parts - X_v and X_t. Each column of X_v represents the visual face image, and each column of X_t represents the thermal face image. Using the dictionary learning, X is factorised into $\hat{D}$ and $\hat{\alpha}$. Colour represents domain specific features and {@, *, #, $, +} represent identity related features. In $\hat{D}$, identity-related atoms are in D_0. Here, the same person's face is represented in two domains, so the person's identity-related features are shared between the two domains. So, D_0 is a common subspace that preserves identity-related features. Each image in X has a common subspace representation code in α_0.

4.3.1.1.1 Common Subspace Learning method1 (CSL1): The inputs to common subspace learning stage are X and the domain labels. At this stage, we learn D, α, D_0 and α_0

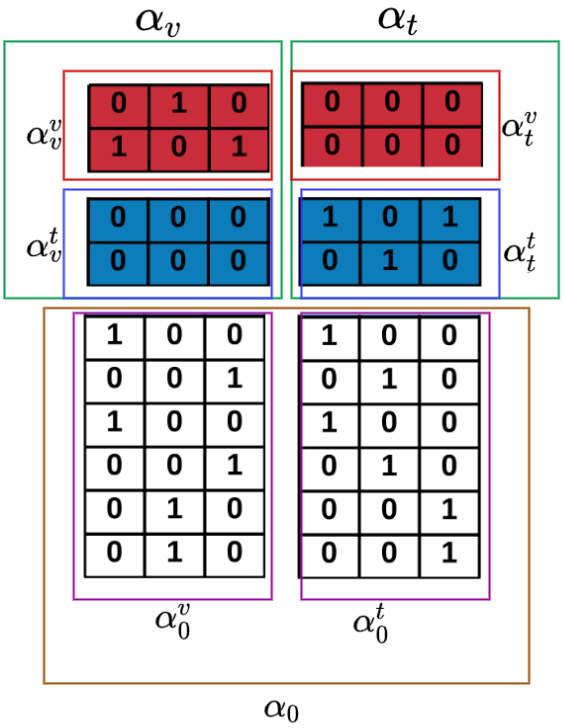


Figure 4.5: Illustration of $\hat{\alpha}$ and notations of its sub parts.

by minimising the following objective function (Equation 4.1).

$$\begin{aligned} J(D, \alpha, D_0, \alpha_0) = \|X - D_0\alpha_0 - D\alpha\|_F^2 + \|X_t - D_0\alpha_0^t - D_t\alpha_t^t\|_F^2 \\ +\|D_v\alpha_v^t\|_F^2 + \|X_v - D_0\alpha_0^v - D_v\alpha_v^v\|_F^2 \\ +\|D_t\alpha_t^v\|_F^2 + \lambda_1\|\alpha\|_1 + \lambda_1\|\alpha_0\|_1 + \lambda_2 f(\hat{\alpha}) \end{aligned} \tag{4.1}$$

Where,

- X is input data with two parts - thermal data X_t and visual data X_v
- D is class specific dictionary
- D_0 is commonality dictionary
- α is class-specific representation code
- α_0 is commonality representation code
- α_0^v is commonality representation code for visual images
- α_0^t is the commonality representation code for thermal images
- α_v is class-specific representation code for visual images

- α_t is class specific representation code for thermal images
- α_t^t and α_v^t are sub matrices of α_t
- α_t^v and α_t^v are sub matrices of α_v
- λ_1, λ_2 are regularization parameters

Figure 4.5 provides an illustration of the representation code $\hat{\alpha}$ and notations of its sub parts. These are essential components for understanding and defining the Equation 4.1 and Equation 4.2.

$$f(\hat{\alpha}) = \|\alpha_v - M_v\|_F^2 + \|\alpha_t - M_t\|_F^2 - \|M_v - M\|_F^2 - \|M_t - M\|_F^2 + \|\alpha_0 - M_0\|_F^2 + \|\alpha\|_F^2 \tag{4.2}$$

where,

- M is the mean of class-specific representation code α
- M_v is the mean of visual class-specific representation code α_v
- M_t is the mean of thermal class-specific representation code α_t
- M_0 is the mean of commonality-specific representation code α_0

The objective function (Equation 4.1) is not jointly convex. However, it is convex w.r.t each of the D, α, D_0, α_0 separately. So one can use an algorithm that alternatively updates each variable by fixing the others.

Updation of D is done by solving Equation 4.3 ($J(D_t)$) and Equation 4.4 ($J(D_v)$). Here $J(D) = J(D_t) + J(D_v)$. Equation 4.3 and 4.4 are solved using online dictionary learning [90].

$$J(D_t) = \underset{D_t}{\operatorname{argmin}} \|X - D_t\alpha_t - D_v\alpha_v\|_F^2 + \|X_t - D_t\alpha_t^t\|_F^2 + \|D_t\alpha_v^t\|_F^2 \tag{4.3}$$

$$J(D_v) = \underset{D_v}{\operatorname{argmin}} \|X - D_v\alpha_v - D_t\alpha_t\|_F^2 + \|X_v - D_v\alpha_v^v\|_F^2 + \|D_v\alpha_t^v\|_F^2 \tag{4.4}$$

Similarly updation of D_0 is done by solving Equation 4.5, which is also solved using online dictionary learning [90].

$$J(D_0) = \underset{D_0}{\operatorname{argmin}} \| \frac{\bar{X} + \tilde{X}}{2} \|_F^2 + \eta \| D_0 \|_F^2 \tag{4.5}$$

where η is learing rate,

$$\bar{X} = X - D\alpha$$

$$\tilde{X} = X_v - D_v \alpha_v^v + X_t - D_t \alpha_t^t$$

Updation of α is done by minimizing both α_t and α_v and is done by minimizing Equations 4.6 and 4.7.

$$J(\alpha_t) = \underset{\alpha_t}{\operatorname{argmin}} \| X_t - D\alpha_t \|_F^2 + \| X_t - D_t \alpha_t^t \| + \| D_v \alpha_t^v \|_F^2 + \| \alpha_t - M_t \|_F^2 \tag{4.6}$$

$$J(\alpha_v) = \underset{\alpha_v}{\operatorname{argmin}} \| X_v - D\alpha_v \|_F^2 + \| X_v - D_v \alpha_v^v \| + \| D_t \alpha_v^t \|_F^2 + \| \alpha_v - M_v \|_F^2 \tag{4.7}$$

Both the equations 4.6 and 4.7 are solved using the Iterative Projective Method [118].

Updation of the α_0 is done by minimizing the equation 4.8.

$$J(\alpha_0) = \underset{\alpha_0}{\operatorname{argmin}} \| \frac{\bar{X} - \tilde{X}}{2} - D_0 \alpha_0 \|_F^2 + \frac{\lambda}{2} \| \alpha_0 - M_0 \|_F^0 + \lambda_1 \| \alpha_0 \|_F^2 \tag{4.8}$$

The updation of α, α_0, and D can be done by solving above equations. LRSDL [135] also optimises in a similar way except for D_0. In LRSDL, it tries to put in a low rank, whereas, in our problem, D_0 is in a higher rank as the identity-related atoms are in D_0. So the updation of D_0 is done using online dictionary learning [90].

4.3.1.1.2 Common Subspace Learning method2 (CSL2): The main difference between CSL1 and CSL2 is that CSL1 doesn't use any identity-specific labels but uses only domain-related labels. CSL2 uses identity-related features and domain-related la-

bels too. CSL2 has three stages. The first and third stages are similar to that of CSL1, whereas the Stage-2 of CSL2 uses only the identity-related labels. The main idea of CSL2 is to learn more refined domain-specific features, which are learned by using identity-related labels. Stage-1 and Stage-3 in CSL2 aim to get the dictionary-aligned common subspace representation code for all images in X.

$$\begin{aligned} J(D, \alpha, D_0, \alpha_0) = \|X^l - D_0\alpha_0 - D\alpha\|_F^2 + \|X_t^l - D_0\alpha_0^t - D_t\alpha_t^t\|_F^2 \\ +\|D_v\alpha_v^t\|_F^2 + \lambda_1\|\alpha\|_1 + \lambda_1\|\alpha_0\|_1 + \|X_v^l - D_0\alpha_0^v - D_v\alpha_v^v\|_F^2 \\ +\|D_t\alpha_t^v\|_F^2 + \sum_{i=1}^{C}\|\alpha_0^{v/i} - \alpha_0^{t/i}\|_F^2 + \lambda_2 f(\hat{\alpha}) \end{aligned} \tag{4.9}$$

Where,

- X^l, X_t^l and X_v^l are the identity labelled training data
- D is class specific dictionary
- D_0 is commonality dictionary
- α is class-specific representation code
- α_0 is commonality representation code
- α_0^v is commonality representation code for visual images
- α_0^t is commonality representation code for thermal images
- α_v is class specific representation code for visual images
- α_t is class specific representation code for thermal images
- α_t^t and α_v^t are sub matrices of α_t
- α_t^v and α_t^v are sub matrices of α_v
- $\alpha_0^{v/i}$ is the commonality representation code for the visual image of class i, and similarly
- $\alpha_0^{t/i}$ is the commonality representation code for the thermal image of class i
- λ_1, λ_2 are regularization parameters

In Stage-1 of CSL2, we estimate D, α, D_0 and α_0 by minimizing the objective function (Equation 4.1). Thereafter the D and D_0 of Stage-2 are initialised with learned D and D_0 of Stage-1, and then we update D, α, D_0 and α_0 by minimising the objective

function (Equation 4.9). The learned D and D_0 of Stage-2 get transferred to Stage-3 of CSL2, i.e. D and D_0 of Stage-3 are initialised with learned D and D_0 of Stage-2. In Stage-3, we fix D and update α, D_0 and α_0 by minimizing the objective function (Equation 4.9). The updation of α and D can be done in a similar way as that of LRSDL [135]. Updation of α_0 varies slightly because of the additional term $\|\alpha_0^{v/i} - \alpha_0^{t/i}\|_F^2$. Updation of α_0 is done by solving Equation 4.10.

$$J(\alpha_0) = \underset{\alpha_0}{\operatorname{argmin}} \|\frac{\bar{X} - \tilde{X}}{2} - D_0\alpha_0\|_F^2 + \frac{\lambda}{2}\|\alpha_0 - M_0\|_F^0 + \lambda_1\|\alpha_0\|_F^2 + \|\alpha_0^{v/i} - \alpha_0^{t/j}\|_F^2 \tag{4.10}$$

where,

$$\bar{X} = X - D\alpha$$

$$\tilde{X} = X_v - D_v\alpha_v^v + X_t - D_t\alpha_t^t$$

Where $\alpha_0^{v/i}$ be the commonality representation code for the visual image of class i, and similarly $\alpha_0^{t/i}$ is the commonality representation code for the thermal image of class i.

Updation of D_0 is done using online dictionary learning [90].

4.3.1.2 Metric Learning

After completion of common subspace learning, we get common subspace representation code α_0 for all data X. This representation code α_0 is not inherently discriminative, and to make it discriminative, we proposed a deep learning-based metric learning architecture.

4.3.1.2.1 Large Scale Metric Learning (LSML) In large-scale metric learning, we estimate the log-likelihood between learned representation codes, and for that, we need to get the distance between representation codes as given below,

$$Dist(\alpha_0^{vi}, \alpha_0^{tj}) = (\alpha_0^{vi} - \alpha_0^{tj})M(\alpha_0^{vi} - \alpha_0^{tj})^T$$

Here M is the Mahalanobis metric, and the vector difference between the representation code is formulated as given below,

$$Diff(\alpha_0^{vi}, \alpha_0^{tj}) = log\left(\frac{\frac{1}{\sqrt{2\pi\|\Sigma_{z_{ij}=0}\|}} exp^{(-1/2(\alpha_0^{vi}-\alpha_0^{tj})^T \Sigma^{-1}_{z_{ij}=0}(\alpha_0^{vi}-\alpha_0^{tj}))}}{\frac{1}{\sqrt{2\pi\|\Sigma_{z_{ij}=1}\|}} exp^{(-1/2(\alpha_0^{vi}-\alpha_0^{tj})^T \Sigma^{-1}_{z_{ij}=1}(\alpha_0^{vi}-\alpha_0^{tj}))}}\right) \tag{4.11}$$

where,

$$\Sigma_{z_{ij}=0} = \sum_{z_{ij}=0} (\alpha_0^{vi} - \alpha_0^{tj})(\alpha_0^{vi} - \alpha_0^{tj})^T$$

$$\Sigma_{z_{ij}=1} = \sum_{z_{ij}=1} (\alpha_0^{vi} - \alpha_0^{tj})(\alpha_0^{vi} - \alpha_0^{tj})^T$$

Here, z is an indicator matrix of size $n_v \times n_t$. The term $z_{ij} = 0$ indicates that α_0^{vi} and α_0^{tj} are of same person's representation codes (positive pair) and, $z_{ij} = 1$ indicates that α_0^{vi} and α_0^{tj} are of different persons representation codes (negative pair). By solving the above equations, we obtain Mahalanobis distance metric M, and the detailed algorithm is in [73].

In our method, the testing procedure is as follows. Given a pair of test images, x_v^i and x_t^j, representing a visual face image and a thermal face image, respectively, we utilise the learned dictionaries from the common subspace learning stage, denoted as D and D_0. We compute the corresponding representation codes, α_0^{vi} and α_0^{tj}, for the visual and thermal images, respectively. We estimate the similarity between the two images by utilising the learned distance metric, M. This similarity estimation allows us to determine whether the two images are similar.

4.3.1.2.2 Deep Metric Learning In this method, we proposed a deep metric learning architecture. It uses two-stream network architecture; one is for the visual domain and the other for the thermal domain. It is a variant of a Siamese network [50], and the main difference with the Siamese network is that it uses the learned common subspace representation in its fully connected layer. In this deep architecture, there are three main stages. They are feature learning layers, feature embedding layers and loss functions. In feature learning, it learns the features using convolution layers. In the feature em-

bedding layer, along with the convolution layer output, it takes the common subspace representation code. This two-stream network is combined with the loss function.

4.3.2 CSL1+LSML (Method-1)

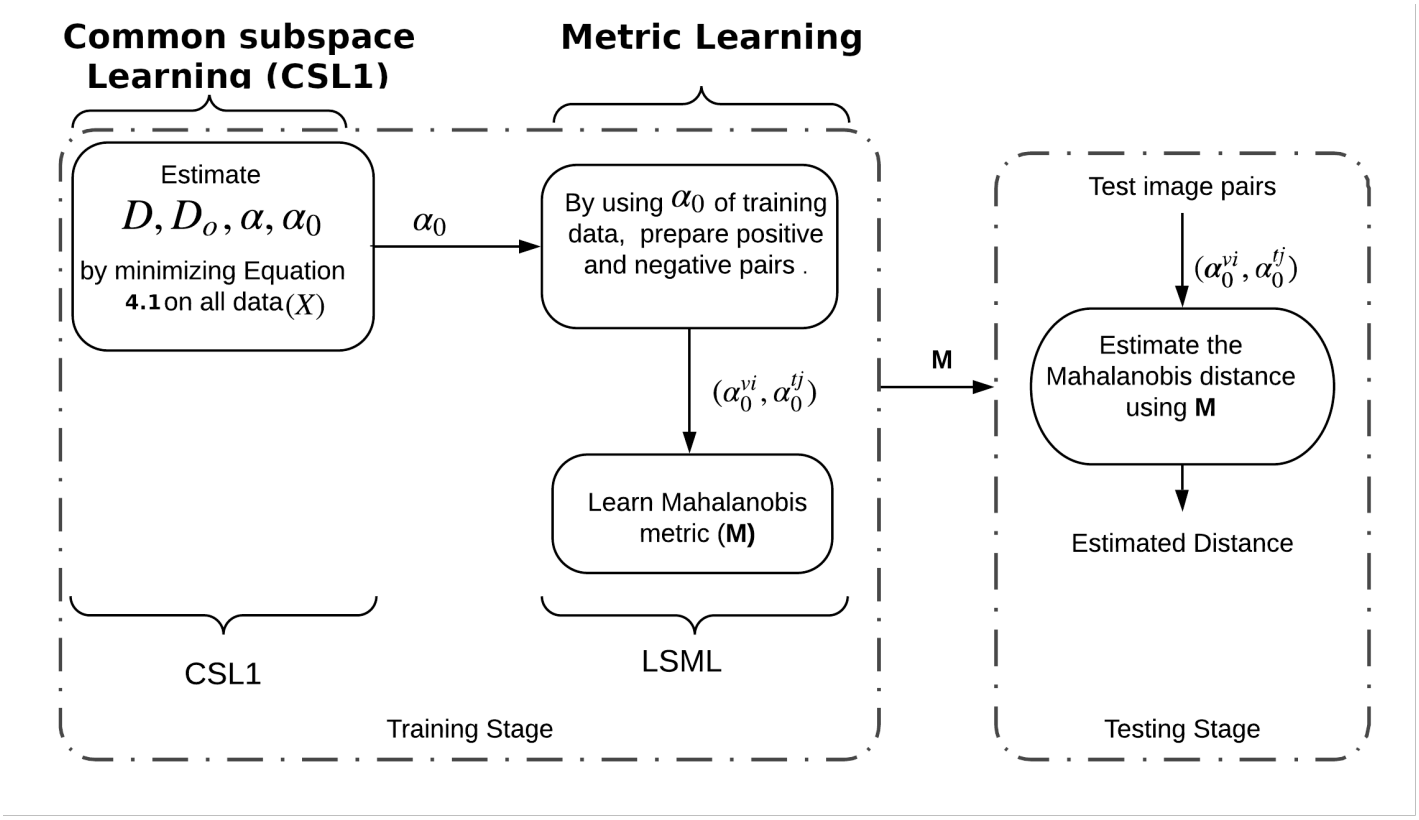


Figure 4.6: Illustration of method-1 (CSL1+LSML). CSL1 is used for common subspace learning, and LSML [73] is used for metric learning.

Figure 4.6 illustrates the proposed method: CSL1+LSML. In the training stage, this method first learns the common subspace using the CSL1 method. In CSL1, it learns the representation code α_0 for all images in X. In the metric learning stage, it learns the metric M using the large scale metric learning (LSML). [73]. In the testing stage, by using learned metric M we estimate the similarity between the representation codes α_0^{vj} and α_0^{ti}. Here α_0^{vj} is a common subspace representation code for the j^{th} visual image. Similarly, i^{th} thermal image common subspace representation code is α_0^{ti}. In this method image to representation code prediction is done using the CSL1 method. The final similarity is predicted with the learned metric of LSML.

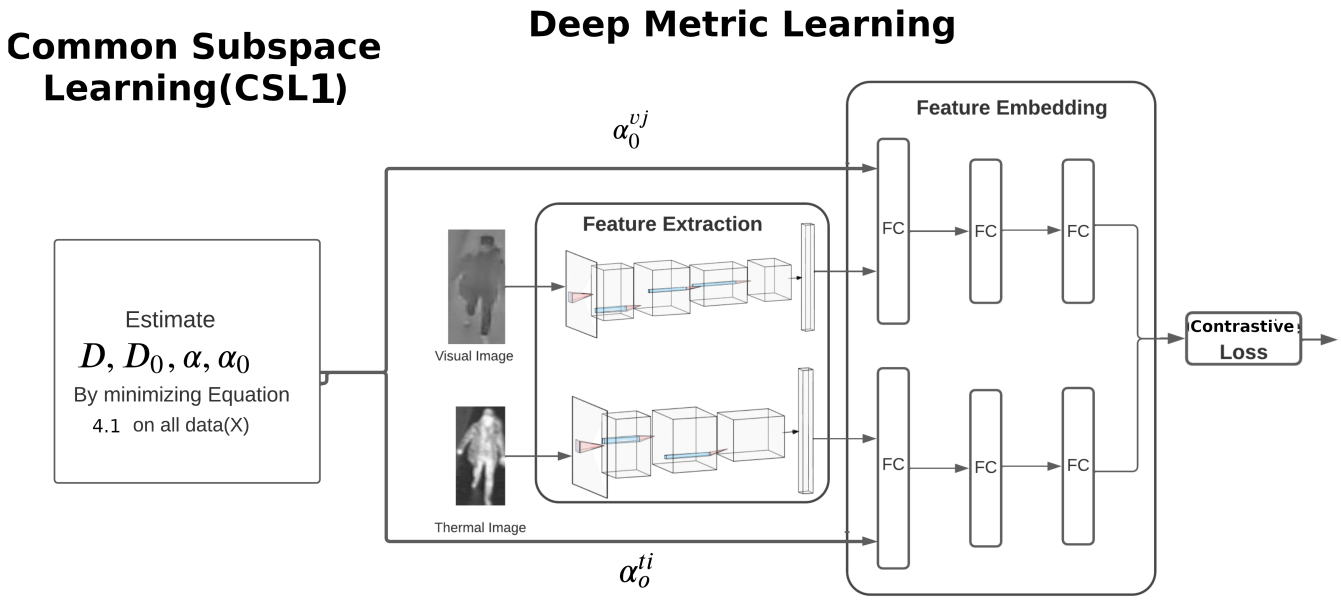


Figure 4.7: Illustration of method-2 (CSL1+DML). CSL1 is used for common subspace learning, and DML is used for metric learning.

4.3.3 CSL1+DML (Method-2)

Figure 4.7 illustrates the proposed method CSL1+DML. In the training stage, this method first learns the common subspace using the CSL1 method, which derives the representation code α_0 for all images in X. Next, in the metric learning stage, the deep metric learning (DML) technique is employed to learn the similarity. During the testing stage, the similarity between the representation codes α_0^{vj} and α_0^{ti} is estimated using the deep metric learning model. Here, α_0^{vj} represents the common subspace representation code for the j-th visual image, while α_0^{ti} represents the common subspace representation code for the i-th thermal image. Similar to the CSL1+LSML method, CSL1+DML utilises the CSL1 method for image-to-representation code prediction. The final similarity is then predicted using the learned DML model, which enhances the discriminative power of the common subspace representation codes.

4.3.4 CSL2+LSML (Method-3)

Figure 4.8 illustrates the proposed method CSL2+LSML. In the training stage, this method first learns the common subspace using the CSL2 method, which captures the

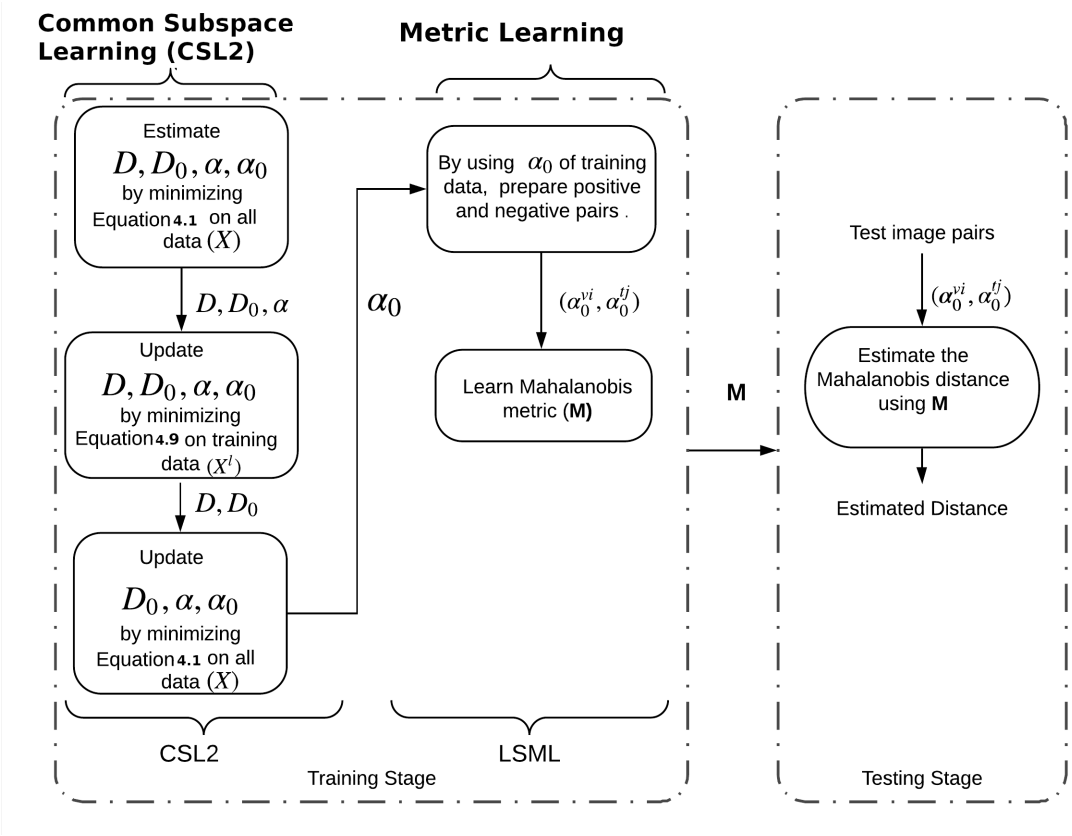


Figure 4.8: Illustration of method-3 (CSL2+LSML). CSL2 is used for common subspace learning, and LSML [73] is used for metric learning.

shared representation code α_0 for all images in X. Subsequently, in the metric learning stage, the large-scale metric learning (LSML) algorithm [73] is employed to learn the metric M. During the testing stage, the similarity between the representation codes α_0^{vj} and α_0^{ti} is estimated using the learned metric M. Here, α_0^{vj} represents the common subspace representation code for the j-th visual image, while α_0^{ti} represents the common subspace representation code for the i-th thermal image. CSL2+LSML utilises the CSL2 method for image-to-representation code prediction. The final similarity is predicted using the learned metric from LSML, which enhances the discriminative power of the common subspace representation codes.

4.3.5 CSL2+DML (Method-4)

Figure 4.9 illustrates the proposed method CSL2+DML. In the training stage, this method first learns the common subspace using the CSL2 method, which obtains the representation code α_0 for all images in X. Next, in the metric learning stage, the deep metric learning (DML) technique is employed to learn the similarity using the learned

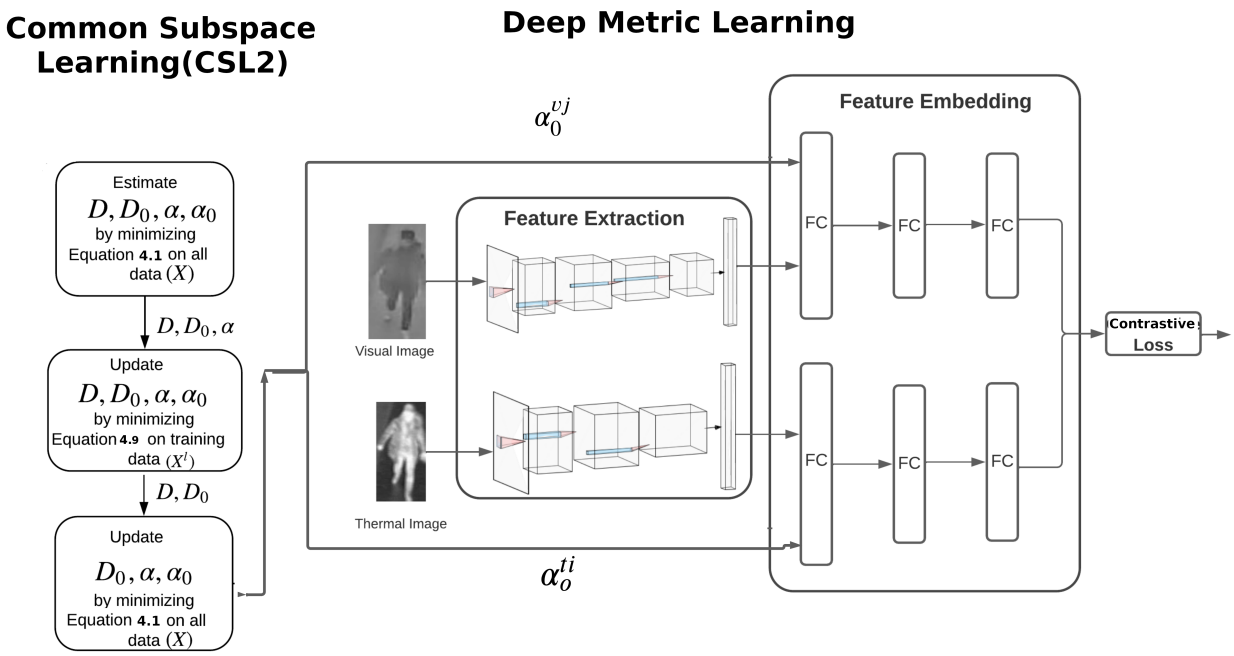


Figure 4.9: Illustration of method-4 (CSL2+DML). CSL2 is used for common subspace learning, and DML is used for metric learning.

DML model. During the testing stage, the similarity between the representation codes α_0^{vj} and α_0^{ti} is estimated using the learned metric M. Here, α_0^{vj} represents the common subspace representation code for the j-th visual image, while α_0^{ti} represents the common subspace representation code for the i-th thermal image. Similar to the CSL2+LSML method, CSL2+DML utilises the CSL2 method for image-to-representation code prediction. The final similarity is predicted using the learned DML model, which improves the discriminative power of the common subspace representation codes.

4.4 Experimental Setup and Results

In common subspace learning training, dictionary size plays an important role. We need to select the size so that the size of the common dictionary (k_0) is greater than that of the domain-specific dictionary (k). The reason behind this is that domain-specific dictionary atoms contribute more to representation but not to discrimination. Redundancy between atoms is high, so the rank of the domain-specific dictionary is low. We experimented with our method with different dictionary sizes and learning rates.

Table 4.1 gives the parameter values for which we got the best results. For CSL1, dictionary size ($\hat{k}$) is 340 (k_t + k_v + k_0) and for CSL2, the dictionary size ($\hat{k}$) is 460 (k_t

		Dictionary Sizes				Regularization Parameters		
Data-Set	**Method**	k_t	k_v	k_0	$\hat{k}$	λ_1	λ_2	λ_3
RGB-D-T	CSL1	20	20	300	340	0.001	0.001	NA
	CSL2	30	30	400	460	0.001	0.001	0.002
RegDB	CSL1	70	70	600	740	0.001	0.001	NA
	CSL2	120	120	800	940	0.001	0.001	0.002

Table 4.1: Parameter values (through which the best results are achieved) used in the proposed method

+ k_v + k_0). The learning rate in metric learning after parameter tuning is set to 0.1 for both methods. To implement CSL1 and CSL2, we have used the discriminating dictionary learning library toolbox [134, 135]. For Deep metric learning, we have used Keras library [24] and TensorFlow library [1]. In this two-stream architecture[69], four convolution layers followed by pooling layers for the visual domain and three convolution layers followed by pooling layers are used for the thermal domain. And then, both the streams have three different fully connected layers followed by the final decision node. For this, we have used contrastive loss.

Dataset Name	Number of Subjects	Number of Visual images per subject	Number of Thermal images per subject	Total number of Visual images	Total number of Thermal images
RegDB[98]	412	10	10	4,120	4,120
RGB-D-T[99]	51	300	300	15,300	15,300

Table 4.2: Details of datasets

Table 4.2 presents details of two datasets: RGB-D-T [99] and RegDB [98]. The RGB-D-T dataset comprises images of 51 individuals captured in three different modalities: visual, depth, and thermal. Each person has 300 images in each modality, resulting in a total of 15,300 (51$\times$300) thermal images and 15,300 (51$\times$300) visual images from this dataset. The visual images have a resolution of 640$\times$480, while the thermal images have a resolution of 384$\times$288. The RegDB dataset contains images of 412 individuals captured in two domains: visual and thermal. Each domain consists of 4,120 images (412$\times$10). Specifically, there are 10 visual images and 10 thermal images for each person in the dataset."

All the thermal images are obtained in the long-wave infrared region. We divide the RGB-D-T dataset into two parts, which consist of 51 different person images. In the first part, we use 26 persons' thermal-visual image pairs for training; in the second part, 25 persons' thermal-visual image pairs are used for testing. The training set and testing set contains different person images, and no person is common between the training set and testing set. Similarly, we split the RegDB dataset into two halves. We employ

206 thermal-visual picture pairings for training and 206 thermal-visual image pairs for testing.

To evaluate the performance of the methods, we followed the protocol used in [159]. The gallery set consists of images from the visual modality, while the probe set comprises images from the thermal modality. We measured the performance of the method using the standard k-rank accuracy and mean average precision (*mAP*).

In our experiment, we divided the dataset into three sets: the train set, validation set, and test set. For the RGB-D-T dataset, which includes 51 different individuals, we used 23 persons' thermal-visual image pairs (23×300=6,900 image pairs) for training, 3 persons' thermal-visual image pairs (3×300=900) for validation, and 25 persons' thermal-visual image pairs (25×300=7,500) for testing. These sets contain images of different individuals, with no overlap between them.

Similarly, for the RegDB dataset, which comprises images of 412 individuals, we used 185 persons' thermal-visual image pairs (185×10=1,850 image pairs) for training, 21 persons' thermal-visual image pairs (21×10=210) for validation, and 206 persons' thermal-visual image pairs (206×10=2,060) for testing. Again, the training set, validation set, and testing set contain images of different individuals without any common person. To ensure statistically stable results, we repeated the experiment ten times.

Data sets	**RGB-D-T**				**RegDB**			
Methods	Rank=1	Rank=5	Rank=20	*mAP*	Rank=1	Rank=5	Rank=20	*mAP*
PLS-DA[51]	33.19	40.98	58.76	25.58	14.16	25.87	41.48	16.44
TONE[159]	36.91	42.12	58.77	29.85	15.77	28.25	45.67	18.75
TONE+HCML[159]	38.49	49.11	63.83	31.53	23.47	36.84	57.53	23.89
BDTR[162]	41.36	62.93	72.82	36.82	34.33	49.09	67.92	33.10
CSL1+LSML(Ours)	35.61	41.59	54.64	28.31	14.96	26.95	44.65	18.44
CSL2+LSML(Ours)	37.92	48.97	61.08	31.24	15.81	28.67	46.21	18.80
CSL1+DML (Ours)	38.25	49.84	64.66	32.11	27.47	42.0	59.66	26.85
CSL2+DML (Ours)	**43.84**	**64.68**	**74.15**	**37.02**	**34.02**	**51.23**	**68.31**	**33.76**

Table 4.3: Thermal to visual cross domain face recognition results

The experimental results on RGB-D-T and RegDB dataset are shown in Table 4.3. Column-1 shows the methods with which we have experimented, whereas the second, third and fourth columns give the Rank-1, Rank-5 and Rank-20 accuracy of the RGB-D-T dataset, and similarly sixth, seventh and eighth columns give the Rank-1, Rank-5 and Rank-20 accuracy of RegDB dataset. The sixth and ninth columns are the *mAP* value of RGB-D-T and RegDB datasets. Higher the value of *mAP*, the better the performance of

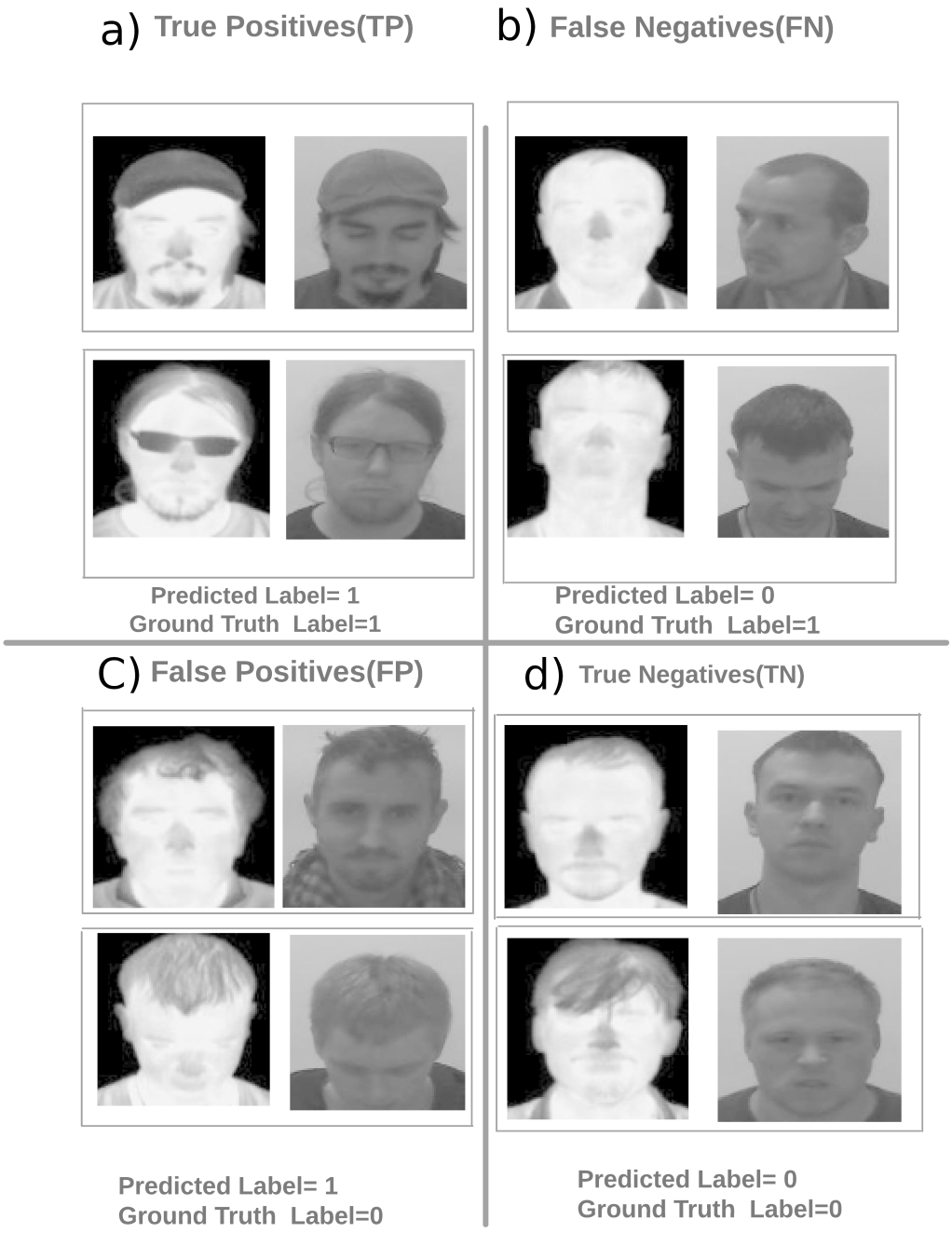


Figure 4.10: Sample thermal-visual image pairs (from RGB-D-T data-set), predicted output labels using method-4 (CSL2+DML), and their ground truth labels

the method. Our proposed methods are compared with the baseline algorithms - PLS-DA [51], TONE[159] and BDTR[162]. Observing the results, we say that CSL2+DML (our method) performs better.

Figure 4.10 illustrates the sample output predictions of method-2 (CSL2+DML). In this context, label = 1 refers to a thermal-visual image pair of the same person, while label = 0 indicates a thermal-visual image pair of different individuals. Figure 4.10 a)

shows the true positives, where both the predicted label and the ground truth label are 1. Figure 4.10 b) displays the false negatives, where the predicted label is 0 and the ground truth label is 1. Figure 4.10 c) illustrates the false positives, where the predicted label is 1 and the ground truth label is 0. Lastly, Figure 4.10 d) depicts the true negatives, where both the predicted label and the ground truth label are 0.

4.5 Summary

We have presented a two-stage cross-domain (thermal to visual) face recognition method based on dictionary learning. Specifically, we have introduced four methods: CSL1+LSML, CSL2+LSML, CSL1+DML, and CSL2+DML. These four methods consist of two variants of common subspace learning methods and two variants of metric learning methods.Common subspace learning is performed to extract identity-related features by removing domain-specific features from the images. We project both domain images onto a common subspace, representing the face images with representation codes. To enhance the discrimination in the representation codes, we utilize metric learning. In the second stage, we employ two variants of metric learning methods: large-scale metric learning and deep metric learning. These methods are applied to improve the discriminative power of the representation codes.

We evaluated the performance of these methods on two datasets: RGB-D-T and RegDB. The test set size for the RGB-D-T dataset consists of 7,500 thermal-visual image pairs from 25 individuals, while the test set size for the RegDB dataset comprises 2,060 thermal-visual image pairs from 206 individuals. Remarkably, the CSL2+DML method outperforms the others, even when there are no common individuals between the training and testing sets.

CHAPTER 5

Thermal to visual face recognition using collaborative metric learning

In the previous chapter, we discussed thermal to visual face recognition using dictionary learning. This chapter discusses the collaborative metric learning (CML) based method. Both methods work on open-set recognition. The collaborative metric learning-based method is more generalised than the dictionary learning method. The CML method learns the more generalised model with less amount of training data.

5.1 Introduction

In this chapter, we propose a collaborative metric learning method using matrix factorization-based collaborative filtering using maximum margin matrix factorization. Figure 5.1 illustrates the proposed method. It takes the visual domain and thermal domain face images as input and returns 'Yes' if they are of the same person's face; otherwise, it returns 'No'.

We first discuss metric learning, collaborative filtering, and the relationship between the two in order to better understand the proposed collaborative metric learning method.

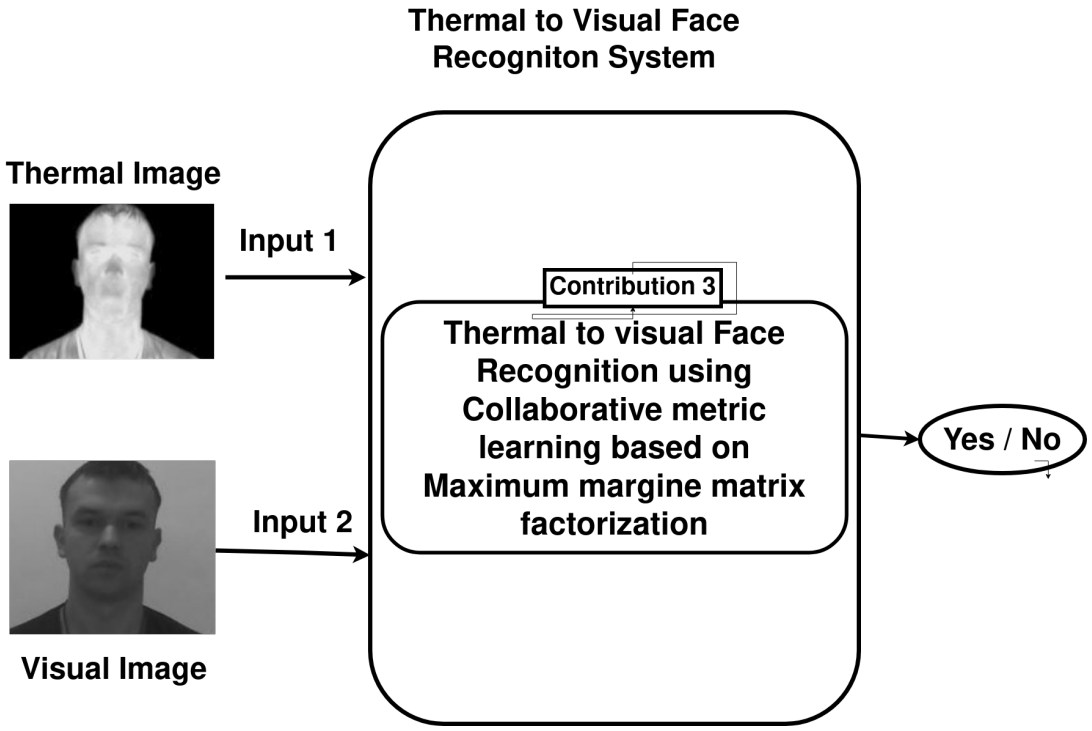


Figure 5.1: Thermal to visual face recognition method using collaborative metric learning based on maximum margin matrix factorization

5.1.1 Metric Learning

Metric learning-based methods are one of the major categories of projection-based techniques. Metric learning seeks to find the similarity between the vectors. Any generalized metric $d(x_1, x_2)$ have the following properties -

- Non-negative similarity $d(x_1, x_2) \geq 0$.
- Identity of indiscernibles $d(x_1, x_2) = 0$ if and only if $x_1 = x_2$
- Symmetric $d(x_1, x_2) = d(x_2, x_1)$.
- Triangle inequality $d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$.

Metric learning typically estimates a matrix S that is positive semi-definite $S \succeq 0$. The learned metric satisfies the properties of symmetry and non-negativity because it is positive and semi-definite. The Mahalanobis distance measures the similarity between two vectors.

$$d_S(x_1, x_2) = \sqrt{(x_1 - x_2)^T S (x_1 - x_2)}$$

Since S is positive semi-definite, it can be factorised as $S = A^T A$

$$d_S(x_1, x_2) = \sqrt{(x_1 - x_2)^T A^T A (x_1 - x_2)}$$

$$d_S(x_1, x_2) = \sqrt{(Ax_1 - Ax_2)^T (Ax_1 - Ax_2)}$$

The preceding equations show that the Mahalanobis distance $d_S(x_1, x_2)$ in the current space equals Euclidean distance in the latent space. Finding the appropriate latent space is one of the objectives of metric learning. It determines similarity by employing positive pairs or negative pairs or triplets (anchor sample, positive sample, and negative sample). In metric learning, positive samples are pulled together more closely while negative samples are pushed farther apart.

5.1.2 Collaborative Filtering

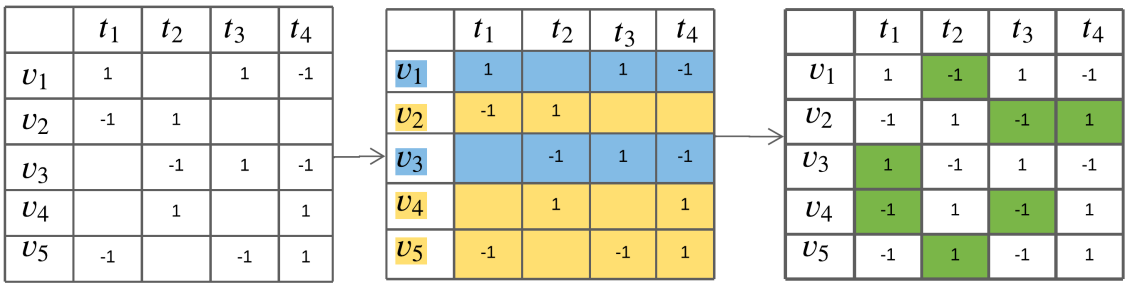


Figure 5.2: Illustration of collaborative filtering. Similarities are shown in the leftmost table. Similarities between the rows are displayed in the centre table. Predicted similarities (green cells) are given in the last table.

Similar to metric learning, collaborative filtering (CF) is employed to determine the similarity among samples. CF serves as a prevalent approach in recommender systems, and its workings are depicted in Figure 5.2. In this method, similarities between two groups of samples are provided. Its objective is to find any missing similarities between the groups. In order to estimate these missing similarities between the groups, we take into account the intra-group similarities. Matrix factorization stands as a highly popular collaborative filtering technique. This method entails projecting the group samples onto a shared subspace while maintaining the provided similarities. Notably, this process inherently preserves the intra-group similarities, thereby contributing to the effectiveness

of the approach.

$$Minimize\mathcal{J} = \sum_{(i,j)\in\omega} \mathcal{L}(S_{ij}, v_i, t_j) + \mathcal{R}(V, T)$$

Here $\mathcal{L}(S_{ij}, v_i, t_j)$ the loss function that operates on the similarity matrix S and the latent factor matrices V and T. It quantifies the discrepancy between the predicted similarities and the actual similarities. Additionally, $\mathcal{R}(V, T)$ represents a regularisation term applied to the latent factor matrices, which aids in controlling overfitting and promoting generalisation. In this context, ω denotes a set of known similarities that are incorporated into the loss function and regularisation term during the collaborative filtering process.

5.1.3 Metric learning and Matrix factorization

Both methods share the objective of discovering the latent space while maintaining the provided similarities. In metric learning, this is accomplished through the utilisation of a positive semi-definite matrix. Each sample is allocated a latent space feature in both approaches. At their initial stages, both methods focus on preserving the given similarities. To achieve greater generalisation, the preservation of the triangle inequality becomes essential. This generalisation is facilitated by carefully selecting an appropriate loss function and regularisation technique that effectively guides the learning process. In this work, we have employed a technique called Maximum Margin Matrix Factorization (MMMF) [130] to achieve a more generalised latent space, even with a limited number of training samples. Our approach involves several steps. First, we construct a similarity matrix S using the labels of the training samples. The rows of the similarity matrix represent thermal domain image samples, while the columns represent visual domain image samples. We construct the similarity matrix using the training samples labels so that the same person images get a high similarity, different person images get a low similarity and zero for unknown pairs. Next, we apply MMMF to factorise the similarity matrix S into two matrices, T and V. Matrix T represents the latent features of the thermal domain images, while matrix V represents the latent features of the visual domain images. To perform regression on T and V, we utilise two separate convolutional neural networks (CNNs). The first CNN, CNN_T, takes thermal images as input

and aims to regress the corresponding latent features in T. Similarly, the second CNN, CNN_V, takes visual images as input and attempts to regress the corresponding latent features in V. In essence, CNN_V acts as a mapping function from the visual domain to the latent space, while CNN_T serves as a mapping function from the thermal domain to the latent space. We iteratively apply MMMF on the similarity matrix S, initialising the values of T and V using the outputs of CNN_T and CNN_V, respectively. We repeat this process until we reach a stable state, which is determined using a validation set. This iterative refinement helps to optimise the latent space representation and improve the performance of the overall model. The proposed method has been evaluated in both few-shot learning and zero-shot learning scenarios. In few-shot learning, there are no constraints on prior knowledge and the aim is to leverage this knowledge to rapidly generalise new tasks with a limited number of training samples. In this context, new tasks with limited training samples refer to situations where we have a new person with only a limited number of training image labels. Zero-shot learning, on the other hand, presents an even more challenging scenario where no training image labels are available for a new person. Specifically, we have conducted tests in both one-shot learning and five-shot learning settings within the few-shot learning paradigm. In one-shot learning, there is only one training image label available per person for new individuals. Similarly, in five-shot learning, we have five training image labels per person. Throughout this chapter, we used the terms few-shot learning with, one-shot learning and/or five-shot learning interchangeably. This chapter makes the following main contributions:

- We tackle the problem of thermal to visual cross-domain face recognition by employing collaborative metric learning. Our proposed method focuses on learning from inter-group similarities, enabling the preservation of both inter-group and intra-group similarities during the learning process.
- We introduce collaborative metric learning using maximum margin matrix factorization. By incorporating a maximum margin approach, we achieve a more generalised metric that enhances the recognition performance.
- In few-shot learning settings, our method surpasses the performance of state-of-the-art algorithms for thermal to visual cross-domain face recognition.

The subsequent sections of this chapter are organised as follows: Section 5.2 presents a discussion on related work in the field. Section 5.3 provides a detailed explanation of our proposed approach. In Section 5.4, we present the experimental results obtained from our method. Finally, Section 5.5 concludes the chapter.

5.2 Related Work

In the existing literature, a significant amount of research has focused on single-domain re-identification, specifically within the visual domain [49, 57, 78]. These studies can be broadly categorised into two main approaches: feature-based strategies and metric learning-based methods. Feature-based methods aim to learn more generalised features [144, 164] and can be further divided into two subcategories: global-based features and local-based features. Global feature learning considers the entire image and learns features from it [19, 80]. On the other hand, local feature learning involves dividing the image into sub-parts and learning features for each individual part [35, 87]. Metric learning-based methods [49, 122] focus on estimating similarities between images. These techniques employ various approaches to learn a metric or distance function that can effectively quantify the similarity between pairs of images.

In scenarios where there is a scarcity of sufficient training data, the performance of cross-domain face recognition systems tends to deteriorate significantly. To address this challenge, few-shot learning strategies have emerged as a promising approach. Few-shot learning techniques aim to enhance the generalisation and discrimination capabilities of cross-domain face recognition models when faced with limited training data. In the context of few-shot learning for cross-domain face recognition, several methods have been proposed. One approach is the introduction of a dual distance metric learning technique, which was presented in [28]. Another approach involves learning color invariant features using metric learning, as demonstrated in [7]. Deep feature learning-based methods are proposed in [129, 150]. Additionally, in [174], a transferable local relative distance comparison method is utilised for few-shot learning-based face recognition.

In the existing literature, cross-domain face recognition has been extensively studied across various modalities, including NIR-visual, 3D-2D, low-resolution-high-resolution, and thermal-visual images. This field is often referred to as heterogeneous face recognition, cross-domain face recognition, or inter-modality face recognition. Cross-domain face recognition approaches can be broadly categorised into three types [103]: feature-based, synthesis-based, and projection-based methods. These approaches aim to identify and extract invariant traits that can be used for recognition across different ap-

plication domains. Feature-based techniques [37] focus on selecting and extracting domain-invariant features. Shuowen *et al.*[51] introduced a method called partial least square discriminant analysis (PLS-DA), which matches manually crafted features such as SIFT, HOG, and LBP from both visual and thermal domain images. Sarfraz *et al.*[120] proposed a deep feature matching-based method. Synthesis-based methods [166] aim to synthesise data in other domains to bridge the gap between different modalities. Generative adversarial networks (GANs) are commonly utilised in synthesis-based approaches [145, 166]. GANs enable the generation of synthetic images in one domain that closely resemble images from another domain, thereby facilitating recognition across different domains. Projection-based methods are widely employed in cross-domain face recognition. These methods involve projecting both domain images into a shared subspace using projection-based techniques [116], where they become more comparable than in their original spaces. Projection-based methods can be categorised into two main categories: dictionary learning and metric learning. In the dictionary learning approach, Reale *et al.*[114] proposed a coupled dictionary learning method. This method involves learning separate dictionaries for each domain, enabling the extraction of discriminative features. On the other hand, deep metric learning methods have been developed specifically for thermal to visual cross-domain face recognition. These methods leverage deep learning architectures and introduce various novel loss functions to enhance the discriminative power of the learned features. For instance, Jambigi *et al.*[59] proposed a maximum mean discrepancy-based loss function.

In the context of cross-domain face recognition, different loss functions have been explored to improve the performance. Some examples include the contrastive loss used in [159], the dual constrained top-ranking loss used in [162], and the intra-modality weighted-part aggregation loss used in [161]. Deep metric learning for cross domain face recognition is successfully generalised by combining the verification loss (such as triplet loss or siamese loss) and ID loss (softmax layer loss).

There are some similarities between our proposed method and cross-domain deep metric learning methods [116, 120]. In deep metric learning approaches, both the identity loss and verification loss are minimised to learn discriminative representations. Similarly, in our method, we minimise the identity loss and verification loss through the use of MMMF. Additionally, we also achieve generalisation by leveraging the maximum margin, which promotes better separation and discrimination discrimination be-

tween different face classes in the latent space.

Mang *et al*.[158, 161] proposed deep metric learning methods for cross-domain face recognition using a two-stream network architecture. In their work, they introduced a typical loss function that combines verification loss and ID loss. The inclusion of the ID loss allows for generalisation of the learned metric. Similarly, our method, which utilises collaborative filtering, inherently incorporates identity information.

Our proposed method facilitates the learning of a more generalised latent space for two primary reasons. Firstly, collaborative filtering inherently preserves both inter-group and intra-group similarities simultaneously, making it aware of the underlying identity information. Secondly, our approach employs Maximum Margin Matrix Factorization (MMMF) with a hinge loss function that maximises the margin. By maximising the margin, we increase the likelihood of preserving the triangle inequality, which further enhances the generalisation capability of the learned metric.

5.3 Proposed Collaborative Metric Learning Method

Figure 5.3 provides an illustration of the proposed collaborative Metric Learning Method (CML). This method consists of three main stages:

- Initialization stage
- Feature mapping stage
- Latent space learning stage

In the Initialization stage, we employ Maximum Margin Matrix Factorization (MMMF) to discover the latent space by preserving the labeled similarities. In the Feature mapping stage, we learn a mapping function that transforms the image space into the latent space. This mapping function enables the representation of face images in a domain-invariant and discriminative manner, facilitating accurate recognition across domains. In the Latent space learning stage, we further enhance the generalisation capability of the learned latent space by utilising the maximum margin. By maximising the margin, we encourage a clear separation between different classes in the latent space, leading to improved discrimination performance.

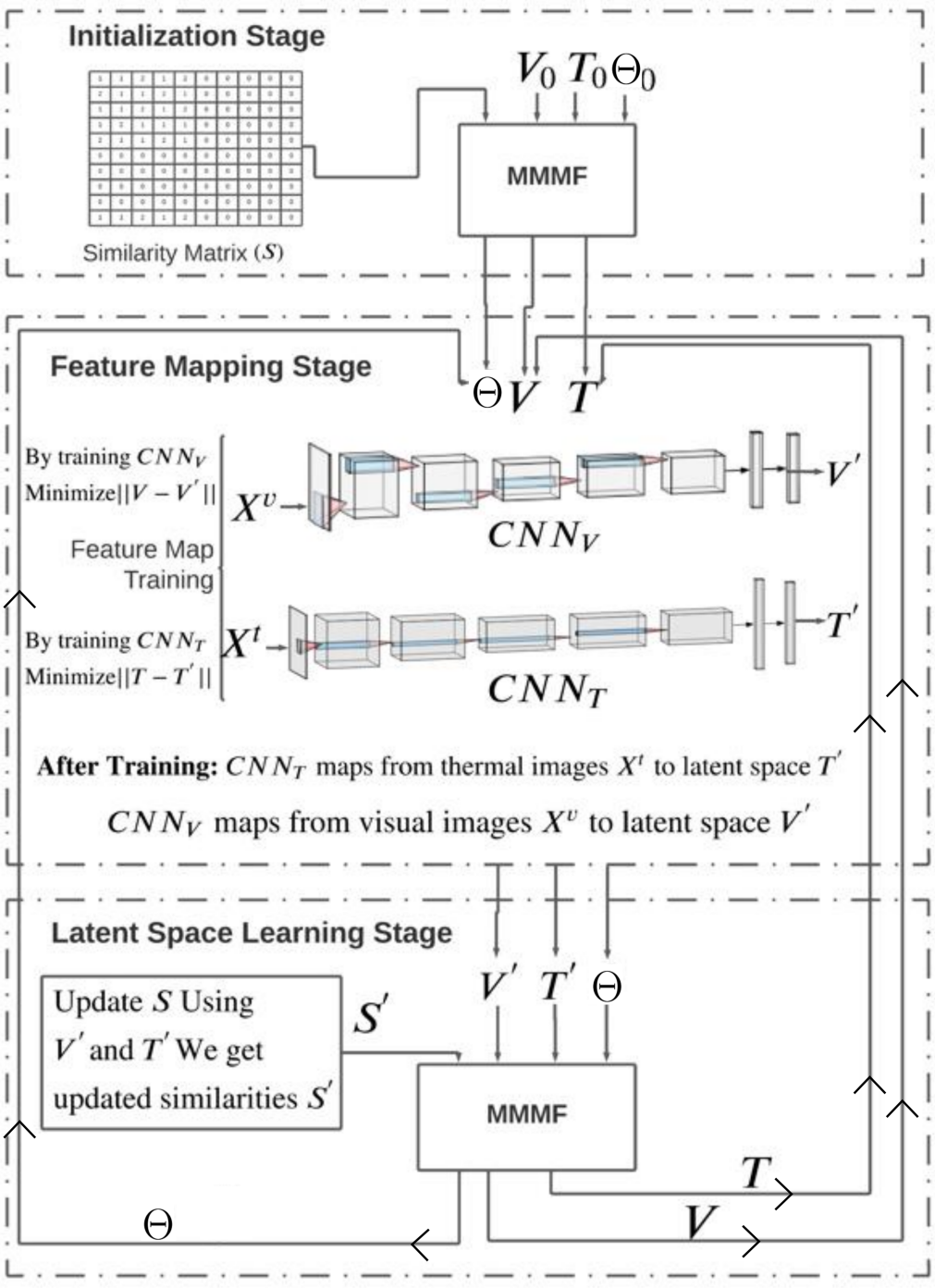


Figure 5.3: Illustration of the proposed method. It consists of mainly three stages 1) Initialization stage, 2) Feature mapping stage, and 3) Latent space learning stage

5.3.1 Initialization stage

In the Initialization stage, the similarity matrix S is constructed from the labels of the training data. The similarity matrix S represents the pairwise similarities between visual and thermal images. Each row in S corresponds to a visual domain image, while each column represents a thermal domain image. The similarity value $S(i, j)$ indicates the degree of similarity between the visual image x_i^v and the thermal image x_j^t, where $x_i^v \in X^v$ and $x_j^t \in X^t$. If the two images belong to the same person, there is a high degree of similarity, and if they are of distinct individuals, the similarity is low. High similarity is represented by 2, low similarity by 1, and unknown similarity by 0 in the matrix S.

In the next step, we apply the Maximum Margin Matrix Factorization (MMMF) technique to the similarity matrix S in order to obtain the visual latent space features V, thermal latent space features T, and threshold values Θ. The MMMF takes V_0, T_0, Θ_0, and S as input and outputs updated latent features V, T, and updated threshold values Θ. It is important to note that the initialization of V_0 and T_0 plays a crucial role in the algorithm, and different initialization methods will be discussed in the subsequent section. The key component of our approach is the use of Maximum Margin Matrix Factorization (MMMF), which plays a crucial role in both the initialization stage and the latent space learning stage.

5.3.1.1 Maximum Margin Matrix Factorization (MMMF)

The key component of our approach is the use of Maximum Margin Matrix Factorization (MMMF). MMMF plays a crucial role in both the initialization stage and latent space learning stage. The main objective of MMMF is to minimise the following objective function $\mathcal{J}(V, T, \Theta)$ (Eq. 5.1) by determining the appropriate latent factor matrices V, T, and thresholds θ_i for each user i. This optimisation process allows us to learn the most discriminative and informative latent representations for cross-domain face recognition.

$$\mathcal{J}(V, T, \Theta) = \sum_{(i,j)\in\omega} l(\mathcal{F}_{ij}(\theta_i - v_i t_j^T)) + \frac{\lambda}{2}(||V||_F^2 + ||T||_F^2) \tag{5.1}$$

where, $\lambda > 0$ is a regularization parameter. $\mathcal{F}_{ij}$ is mapping function defined as,

$$\mathcal{F}_{ij} = \begin{cases} +1 & \text{if } S_{ij} = 2 \\ -1 & \text{if } S_{ij} = 1 \end{cases}$$

and $l(\cdot)$ is the smoothed hinge-loss function defined as,

$$l(x) = \begin{cases} 0 & \text{if } x \geq 1 \\ \frac{1}{2}(1-x)^2 & \text{if } 0 < x < 1 \\ \frac{1}{2} - x & \text{otherwise} \end{cases}$$

The optimisation function of MMMF can be solved using the gradient descent method by updating V, T, Θ, using Equation 5.2.

$$\begin{aligned} V^{t+1} &= V^t - c\,\frac{\partial \mathcal{J}}{\partial V} \\ T^{t+1} &= T^t - c\,\frac{\partial \mathcal{J}}{\partial T} \\ \Theta^{t+1} &= \Theta^t - c\,\frac{\partial \mathcal{J}}{\partial \Theta} \end{aligned} \tag{5.2}$$

Here c is the learning rate. $\frac{\partial \mathcal{J}}{\partial V}$, $\frac{\partial \mathcal{J}}{\partial T}$, $\frac{\partial \mathcal{J}}{\partial \Theta}$ are the partial derivatives (gradients) of $\mathcal{J}$ w.r.t V, T, Θ and are given below (Equations 5.3, 5.4, 5.5).

$$\frac{\partial \mathcal{J}}{\partial V_{ik}} = \lambda V_{ik} - \sum_{j|ij\in\omega} \mathcal{F}_{ij}.l'(\mathcal{F}_{ij}(\theta_i - V_i T_j^T))T_{jk} \tag{5.3}$$

$$\frac{\partial \mathcal{J}}{\partial T_{jk}} = \lambda T_{jk} - \sum_{i|ij\in\omega} \mathcal{F}_{ij}.l'(\mathcal{F}_{ij}(\theta_i - V_i T_j^T))V_{ik} \tag{5.4}$$

$$\frac{\partial \mathcal{J}}{\partial \Theta_i} = \sum_{j|ij\in\omega} \mathcal{F}_{ij}.l'(\mathcal{F}_{ij}(\theta_i - V_i T_j^T)) \tag{5.5}$$

where,

$$l'(x) = \begin{cases} 0 & \text{if } x \geq 1 \\ x - 1 & \text{if } 0 < x < 1 \\ -1 & \text{otherwise} \end{cases}$$

By applying Maximum Margin Matrix Factorization (MMMF) on the similarity matrix, we collaboratively predict the latent features V, T, and Θ. In this process, we iteratively update the latent features while assuming one feature to be fixed at each step. By minimising the norm of the other latent feature during these updates, we aim to maximise the margin for the prediction. The learned V, T, and Θ are passed to the next stage.

5.3.2 Feature mapping stage

The second stage of the proposed method focuses on learning the mapping functions from visual images to latent space features and from thermal images to latent space features. The latent space features are obtained from either the initialization stage or the latent space learning stage, depending on the progression of the algorithm. In this stage, convolutional neural networks (CNNs) are employed as the mapping functions. For the CNN architecture, we have used the AlexNet[72] architecture with a simple modification: the loss function used here is cosine similarity, as it is a regression network.

Specifically, CNN_v is trained to map visual images X^v to the latent space visual features V. Similarly, CNN_t is trained to map thermal images X^t to the latent space thermal features T. The predicted latent features, denoted as V' and T', are generated by CNN_v and CNN_t, respectively. The training objective of this stage is to minimize the distance between the input latent features and the predicted latent features. Therefore, CNN_v aims to minimize $||V - V'||$, while CNN_t aims to minimize $||T - T'||$.

In this stage, Θ, the predicted latent space features V' and T', are passed to the next stage. It is important to note that Θ remains unchanged and is not updated at this stage. The same Θ from the previous stage is carried forward to the subsequent stage.

5.3.3 Latent space learning stage

In the third stage of the proposed method, the focus is on learning the latent space. The first step involves updating the similarity matrix, denoted as S'. This update is based on the predicted latent space features V', T', and the threshold values Θ. The update only applies to the unknown similarities, which refer to similarities that are not present in the training set.

To update S', the cosine similarity between v_i' and t_i' is computed. If the computed similarity deviates from the corresponding threshold Θ_i, then those similarities are updated in S'. Afterwards, MMMF is applied to S' using the initial values of V', T', and Θ. The latent space features, namely V, T, and Θ, are then updated according to the updated similarities in S'. These updated features are passed to the next stage.

This method involves a loop between stage 2 and stage 3, which continues until a termination condition is met. In this case, the termination condition is to minimize the minimum validation error. Once this condition is satisfied, the method returns the final mapping functions and the updated Θ.

Algorithm 4: Collaborative Metric Learning (CML)

Input : S, X^v, X^t, V_0, T_0
/* Here, S is the similarity matrix, X^v is the set of visual images, X^t is the set of thermal images, V_0 is the set of initial visual features, and T_0 is the set of initial thermal features. */
Output: CNN_V, CNN_T
/* Where, CNN_V is the learnt visual mapping function and CNN_T is the learnt thermal mapping function */

```
// Initialization Stage
```
Initialize Θ_0 with random numbers;
[V, T, Θ] = MMMF(S, V_0, T_0, Θ_0);
CNN_V = $CNN_V.initialize()$;
CNN_T = $CNN_T.initialize()$;
```
// Feature Mapping stage
```
while *Stopping criteria For CMF* **do**
```
    // Training the CNN_V to map from X^v to V
```
    $CNN_V.fit(X^v, V)$;
```
```
    // Training the CNN_T to map from X^t to T
```
    $CNN_T.fit(X^t, T)$;
    **if** *Not met the Stopping criteria For CMF* **then**
```
        // Predicting latent features from the trained
           networks
```
        $V^{'}$ = $CNN_V$.predict($X^v$);
        $T^{'}$ = $CNN_T$.predict($X^t$);
```
        // Latent Space learning Stage
```
        Update $S$ using $V^{'}, T^{'}$ and $\Theta$ and get $S^{'}$;
        [$V, T, \Theta$] = MMMF($S^{'}$, $V^{'}$, $T^{'}$, $\Theta$);
    **end**
**end**
**return** $CNN_V$ and $CNN_T$;

Algorithm 4 outlines the proposed collaborative metric learning method. The algo-
```

rithm takes several inputs, including the similarity matrix S, visual images X^v, thermal images X^t, initial visual features V_0, and initial thermal features T_0. Its aim is to learn the visual mapping function CNN_V and thermal mapping function CNN_T as the outputs.

Algorithm 5: Maximum Margin Matrix Factorization (MMMF)

Input : S, V_0, T_0, Θ_0

/* Here, S is the similarity matrix, V_0 is the set of initial visual features, and T_0 is the set of initial thermal features. */

Output: V, T and Θ

/* Where, V is the learnt visual latent space, and T is the learnt thermal latent space. */

$t \leftarrow 0$;

Initialize: $V^t = V_0$, $T^t = T_0$, $\Theta^t = \Theta_0$, c=0.01;

ω is a set of all existing similarities in S;

while *Stopping criteria For MMMF* **do**

  $\forall\omega$ Calculate $\frac{\partial J}{\partial V^t}$, $\frac{\partial J}{\partial V^t}$ and $\frac{\partial \mathcal{J}}{\partial \Theta^t}$;

  $V^{t+1} \leftarrow V^t - c\frac{\partial J}{\partial V^t}$; // Updation of visual latent space V

  $T^{t+1} \leftarrow T^t - c\frac{\partial J}{\partial T^t}$; // Updation of thermal latent space T

  $\Theta^{t+1} = \Theta^t - c\,\frac{\partial \mathcal{J}}{\partial \Theta^t}$; // Updation of threshold values Θ

  $t \leftarrow t + 1$;

end

$V = V^t$, $T = T^t$ and $\Theta = \Theta^t$;

return V, T and Θ;

Algorithm 5 demonstrates the maximum margin matrix factorization (MMMF). It takes the initial visual features V_0, initial thermal features T_0, initial threshold values Θ_0, and the similarity matrix S as inputs. It outputs the updated visual features V, updated thermal features T, and the updated threshold values Θ.

In our proposed method, based on the initialization of V_0 and T_0, two variants of the proposed methods are introduced: CML and CML+BDTR. For the CML method, the initialization of V_0 and T_0 is performed using the last fully connected layer of a pre-trained CNN. Specifically, V_0 is initialized with the last fully connected layer of a CNN trained on visual training data, while T_0 is initialized with the last fully connected layer of a CNN trained on thermal training data.

In the CML+BDTR method, a variant of our approach, the initialization of V_0 and T_0 is slightly different. In this case, V_0 and T_0 are initialized using the last fully connected layer of the BDTR network, which is a two-stream network architecture combining

identity loss and verification loss

5.4 Experimental Setup and Results

The proposed method is implemented using both Matlab and Keras. The Keras library [24] is utilised for developing the mapping functions. Specifically, a seven-layer convolutional neural network is employed for this purpose, comprising four convolution layers and three fully connected layers. The loss function used is cosine similarity. The maximum margin matrix factorization (MMMF) is implemented using a Matlab implementation similar to the one described in [130].

We conducted performance evaluations of our proposed method using two datasets: RGB-D-T [99] and RegDB [98]. The RGB-D-T dataset consists of images of 51 individuals captured in three different modalities: visual, depth, and thermal. Each person has 300 images in each modality. In our experiments, we focused on the thermal and visual images, resulting in a total of 15,300 (51×300) thermal images and 15,300 (51×300) visual images from this dataset. The RegDB dataset contains images of 412 individuals captured in two domains: visual and thermal. Each domain consists of 4,120 images (412×10). Specifically, there are 10 visual images and 10 thermal images for each person in the dataset. Further details about the datasets are provided in Table 5.1.

Dataset Name	Number of Subjects	Number of Visual images per subject	Number of Thermal images per subject	Total number of Visual images	Total number of Thermal images
RegDB[98]	412	10	10	4,120	4,120
RGB-D-T[99]	51	300	300	15,300	15,300

Table 5.1: Details of Datasets

To assess the effectiveness of our proposed method, we employed the following evaluation metrics: Rank-1 accuracy, Rank-5 accuracy, and Mean Average Precision (*mAP*). These metrics are well-suited for scenarios where multiple ground truths exist for a single person in the gallery set [173].

5.4.1 Experimental setup

The experiments were conducted using the following methodology. The dataset was divided into three distinct subsets: a train set, a validation set, and a test set. Each of

these subsets was further divided into two sets: a gallery set and a probe set. The gallery set consisted of visual images, while the probe set consisted of thermal images. For dataset preparation, we followed a similar approach as in [158]. One part of the dataset was allocated for testing purposes, similar to [158]. The remaining part was divided into train and validation sets. We evaluated the proposed methods in three different settings: zero-shot learning, one-shot learning, and five-shot learning.

- **Zero-shot learning setup**
  - In the zero-shot learning setup, there were no common individuals among the train, validation, and test sets. The dataset was divided into three sets: the train set, validation set, and test set.
  - For the RGB-D-T dataset, which includes 51 different individuals, we used 23 persons' thermal-visual image pairs (23×300=6,900 image pairs) for training, 3 persons' thermal-visual image pairs (3×300 = 900) for validation, and 25 persons' thermal-visual image pairs (25×300 = 7,500) for testing. These sets contained images of different individuals with no overlap between them.
  - Similarly, for the RegDB dataset, which comprises images of 412 individuals, we used 185 persons' thermal-visual image pairs (185×10 = 1,850 image pairs) for training, 21 persons' thermal-visual image pairs (21×10 = 210) for validation, and 206 persons' thermal-visual image pairs (206×10 = 2,060) for testing. Once again, the training set, validation set, and testing set contained images of different individuals without any common person.

- **One-shot learning setup**
  - In the one-shot learning setup, the training set was divided into two parts: part-1 and part-2. These parts did not contain any common images of individuals. Part-1 was exclusively used for learning prior knowledge, while from part-2, only one image per person was included in the training set. The remaining images of each person were either included in the test or validation sets.
  - For the RGB-D-T dataset, which includes 51 different individuals, we used 23 (part-1)+ 28 (part-2) persons' thermal-visual image pairs (23×300 (part-1) + 28×1 (part-2) = 6,928 image pairs) for training, 3 persons' thermal-visual image pairs (3×299 = 897) for validation, and 25 persons' thermal-visual image pairs (25×299 = 7,475) for testing.
  - Similarly, for the RegDB dataset, which comprises images of 412 individuals, we used 185 (part-1) + 227 (part-2) persons' thermal-visual image pairs (185×10 (part-1) + 227 (part-2)×1 = 2,077 image pairs) for training, 21 persons' thermal-visual image pairs (21×9 = 189) for validation, and 206 persons' thermal-visual image pairs (206×9 = 1,854) for testing.

- **Five-shot learning setup**
  - Similar to the one-shot learning setup, in the five-shot learning setup, only five images per person were included in part-2.

- For the RGB-D-T dataset, which includes 51 different individuals, we used 23 (part-1)+ 28 (part-2) persons' thermal-visual image pairs (23×300 (part-1)+28×5 (part-2) = 7,040 image pairs) for training, 3 persons' thermal-visual image pairs (3×295 = 885) for validation, and 25 persons' thermal-visual image pairs (25×295=7,375) for testing.
- Similarly, for the RegDB dataset, which comprises images of 412 individuals, we used 185 (part-1) + 227 (part-2) persons' thermal-visual image pairs (185×10 (part-1) + 227×5 (part-2) =2,985 image pairs) for training, 21 persons' thermal-visual image pairs (21×5=105) for validation, and 206 persons' thermal-visual image pairs (206×5=1,030) for testing.

In the one-shot learning scenario, the test set and validation set person images are included in the train set, but with the constraint that only one image per person is used for training. There are no limitations on the number of images per person in the train set. Likewise, in the five-shot learning setup, the train set contains the person images from the test set and validation set, with the inclusion of five images per person for training purposes.

We conducted a comparative analysis of our proposed method with several existing methods, including:

- **Tone [159]**: This method utilises a two-stream network architecture and combines identity loss and verification loss.
- **Tone+HCML [159]**: Building upon the Tone architecture, this method further enhances the results by incorporating a metric learning technique called HCML.
- **BDTR [162]**: BDTR employs a two-stream network architecture and combines identity loss and verification loss. For verification loss, the top-ranking loss is used.
- **CSL1+LSML [36]**: In this method, an unsupervised common subspace is learned using dictionary learning and further optimised with large-scale metric learning.
- **CSL2+LSML [36]**: Similar to CSL1+LSML, this method focuses on learning a supervised common subspace using dictionary learning and large-scale metric learning.
- **MACE [158]**: MACE incorporates modality-shareable and modality-specific classifiers through an ensemble learning approach.

Table 5.2 presents the experimental results of thermal-to-visual cross-domain face recognition on the RegDB dataset. It provides the performance for three different learning settings: zero-shot, one-shot, and five-shot learning. The evaluation metrics reported in the table include Rank-1 accuracy, Rank-5 accuracy, and Mean Average Precision (*mAP*). Higher values for these metrics indicate better performance. The structure

Setting	Zero-shot learning			One-shot learning			Five-shot learning		
Methods	**Rank=1**	**Rank=5**	**mAP**	**Rank=1**	**Rank=5**	**mAP**	**Rank=1**	**Rank=5**	**mAP**
TONE [159]	15.27	25.87	16.44	52	66.8	49.6	75.61	82.9	66.8
TONE+HCML [159]	23.47	36.84	23.89	60.1	69.7	55.2	84.6	91.1	77.6
BDTR [162]	34.33	49.09	33.10	71.12	78.46	60.9	86.82	92.69	81.63
CSL1+LSML [36]	14.96	26.95	18.44	46.87	63.93	44.18	76.86	84.68	65.14
CSL2+LSML [36]	15.81	28.67	18.8	59.83	67.24	54.69	79.26	89.63	75.86
MACE [158]	**72.12**	**80.4**	**68.54**	92.72	95.18	86.46	98.85	99.39	94.84
CML (Ours)	56.46	64.92	40.36	89.86	94.67	85.12	97.69	98.94	93.41
CML+BDTR (Ours)	69.24	74.32	61.5	**96.24**	**99.16**	**90.46**	**98.85**	**99.39**	**94.84**

Table 5.2: Thermal to visual cross domain face recognition results on RegDB

Setting	Zero-shot learning			One-shot learning			Five-shot learning		
Methods	**Rank=1**	**Rank=5**	**mAP**	**Rank=1**	**Rank=5**	**mAP**	**Rank=1**	**Rank=5**	**mAP**
TONE [159]	36.91	42.12	29.85	62.68	71.29	58.72	76.89	84.66	67.21
TONE+HCML [159]	38.49	49.11	31.53	64.39	74.82	60.49	82.69	90.65	79.68
BDTR [162]	41.36	62.94	36.32	74.42	80.67	64.51	84.93	91.81	80.84
CSL1+LSML [36]	35.61	41.59	28.31	61.41	68.88	58.17	77.23	85.39	65.31
CSL2+LSML [36]	37.92	48.97	31.24	64.14	71.64	59.67	80.68	89.64	76.15
MACE [158]	**76.65**	**89.46**	**70.37**	90.97	94.18	85.77	92.66	96.52	91.92
CML (Ours)	59.44	68.46	43.62	88.77	93.41	84.62	95.91	98.71	92.68
CML+BDTR (Ours)	71.65	77.89	66.84	**95.66**	**98.73**	**89.77**	**96.63**	**99.15**	**94.47**

Table 5.3: Thermal to visual cross domain face recognition results on RGB-D-T

of the table consists of columns representing the evaluated methods and the corresponding results for zero-shot, one-shot, and five-shot learning in Column-2, Column-3, and Column-4, respectively.

Similarly, Table 5.3 demonstrates the experimental results of thermal-to-visual cross-domain face recognition on the RGB-D-T dataset. This table follows the same structure as Table 5.2

By observing the results, it can be concluded that the MACE method [158] performs better in the zero-shot learning setting on both datasets. However, in the one-shot and five-shot learning settings, the proposed method CML+BDTR outperforms MACE. The BDTR features serve as a good initialization for the proposed method because BDTR is trained using both the identity loss and the verification loss. As a result, the inclusion of BDTR in CML+BDTR leads to improved performance compared to the CML method alone.

In addition to the CML and CML+BDTR methods, we also explored the option of initializing V_0 and T_0 with features obtained from the MACE method. However, we observed no significant improvement in generalization when using MACE features due to the architectural differences between MACE and the feature mapping stage. The underperformance of the proposed methods in the zero-shot learning setting can be attributed

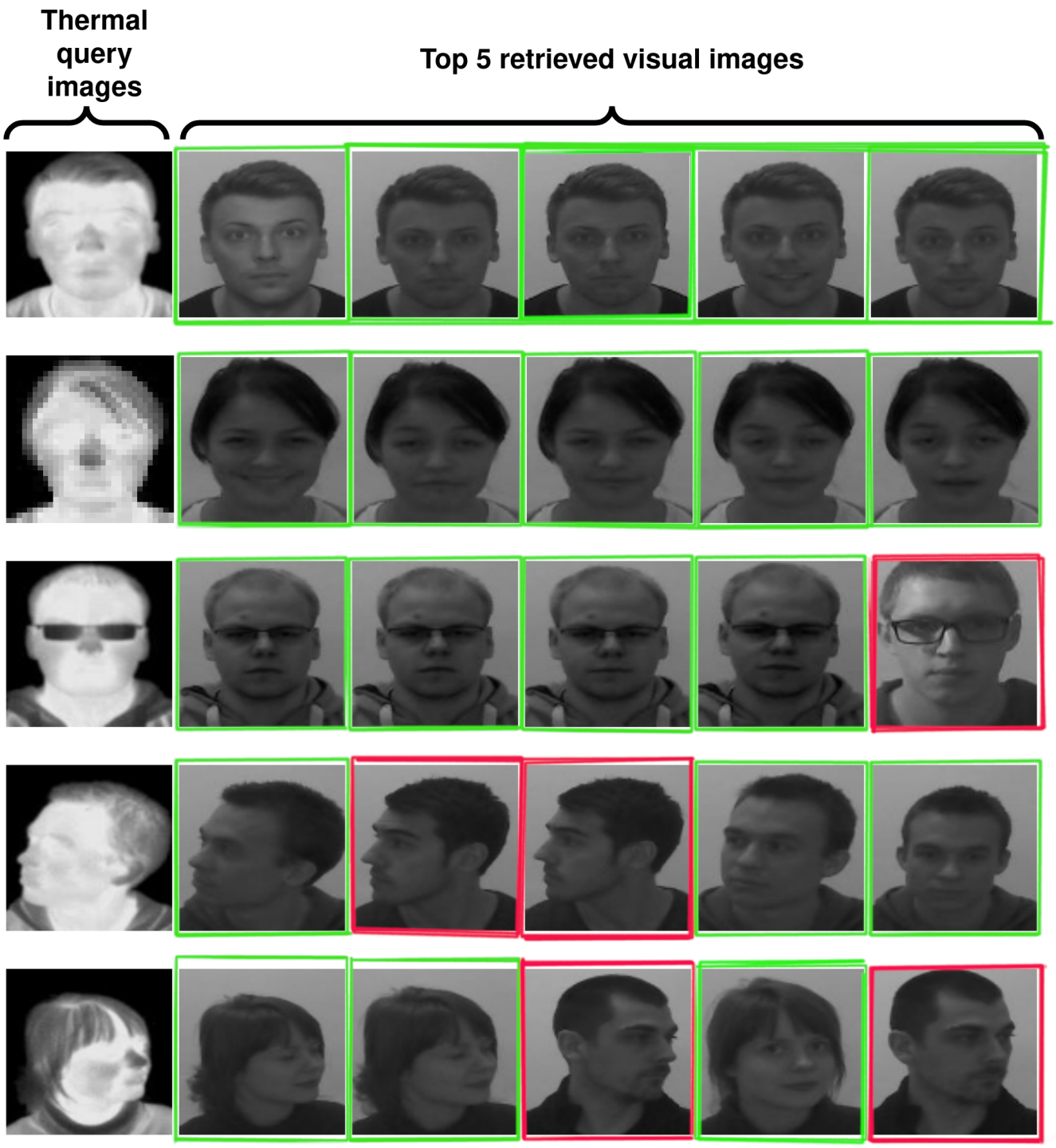


Figure 5.4: Sample results of the proposed method: Five query thermal images (which are in the first column) and their corresponding top-5 retrieved image results from the gallery set (which are in the remaining columns). Correctly retrieved samples are in green boxes, and incorrect matches are in red boxes.

to the cold-start problem associated with the matrix factorization used in our method (MMMF). The cold-start problem arises when there are very few available similarities for a particular image. In such cases, the similarity matrix may not be effectively updated, leading to lower accuracies.

However, in the one-shot and five-shot learning settings, the proposed methods show better accuracies. This improvement can be attributed to the fact that in these settings, there is a sufficient number of labeled samples available for each class. The mapping function, along with the updated similarity matrix, can effectively utilize this limited labeled data, leading to improved performance.

Overall, the proposed method is advantageous when dealing with a limited number of labeled samples per class, as it leverages collaborative metric learning and the mapping function to effectively learn from the available data and improve the accuracy in the one-shot and five-shot learning scenarios.

In Figure 5.4, sample results of CML+BDTR are depicted. The figure showcases five query thermal images in the first column, along with their corresponding top-5 retrieved image results from the gallery set in the remaining columns. Correctly retrieved samples are marked with green boxes, while incorrect matches are highlighted with red boxes. These results highlight the influence of rotation and pose variations on the accuracy of the retrieval process.

5.4.2 Flexibility of proposed method

To evaluate the flexibility of the proposed method, we have conducted experiments on visual-to-thermal cross-domain face recognition. The same training set, validation set, and test set were used, with the only difference being that the query images were visual images while the gallery images were thermal images.

Setting	**Zero-shot learning**			**One-shot learning**			**Five-shot learning**		
Methods	**Rank=1**	**Rank=5**	**mAP**	**Rank=1**	**Rank=5**	**mAP**	**Rank=1**	**Rank=5**	**mAP**
TONE [159]	15.95	26.76	16.94	52.89	67.54	49.85	76.12	83.26	67.05
TONE+HCML [159]	25.6	37.78	24.04	60.72	70.08	60.94	84.95	91.65	77.95
BDTR [162]	33.78	49.69	33.71	71.91	79.36	61.45	87.12	92.98	81.76
CSL1+LSML [36]	15.23	27.35	18.74	47.51	64.92	44.96	77.65	85.05	65.61
CSL2+LSML [36]	16.22	29.48	19.1	60.12	67.95	54.84	79.72	90.04	75.97
MACE [158]	**72.37**	**81.14**	**69.09**	93.01	95.85	86.95	94.65	95.85	92.14
CML (Ours)	57.23	65.56	40.71	90.06	94.92	85.8	97.9	99.01	93.47
CML+BDTR (Ours)	69.92	76.6	61.92	**96.37**	**99.18**	**91.05**	**98.89**	**99.45**	**94.5**

Table 5.4: Visual to thermal cross domain face recognition results on RegDB

Table 5.4 showcases the experimental results of visual-to-thermal cross-domain face recognition on the RegDB dataset. It presents the performance for zero-shot, one-shot, and five-shot learning scenarios, measured through Rank-1 accuracy, Rank-5 accuracy,

Setting	Zero-shot learning			One-shot learning			Five-shot learning		
Methods	**Rank=1**	**Rank=5**	**mAP**	**Rank=1**	**Rank=5**	**mAP**	**Rank=1**	**Rank=5**	**mAP**
TONE [159]	37.62	43.12	30.15	63.32	72.16	59.06	77.26	85.58	68.12
TONE+HCML [159]	41.14	49.94	31.97	64.91	75.35	60.9	83.02	90.96	79.88
BDTR [162]	42.1	63.45	37.21	75.16	81.75	64.96	85.86	92.43	81.25
CSL1+LSML [36]	36.34	42.12	28.96	62.26	69.78	58.89	77.34	85.9	66.84
CSL2+LSML [36]	38.22	49.21	31.63	64.76	72.52	60.62	81.45	90.42	77.11
MACE [158]	**76.92**	**90.06**	**70.78**	91.41	94.89	85.95	93.1	96.98	92.19.
CML (Ours)	59.94	68.92	43.97	89.42	93.96	84.9	96.41	98.89	92.9
CML+BDTR (Ours)	72.16	78.33	67.27	**96.02**	**98.87**	**89.94**	**96.81**	**99.28**	**94.68**

Table 5.5: Visual to thermal cross domain face recognition results on RGB-D-T

and Mean Average Precision (*mAP*). The table's structure includes columns representing the evaluated methods, with the corresponding results for zero-shot, one-shot, and five-shot learning presented in Column-2, Column-3, and Column-4, respectively.

Similarly, Table 5.5 displays the experimental results of visual-to-thermal cross-domain face recognition on the RGB-D-T dataset. The table follows the same structure as Table 5.4.

Upon observing the results, similar to thermal-to-visual face recognition, we find that the MACE method [158] performs better in the zero-shot learning setting on both datasets. However, in the one-shot and five-shot learning settings, the proposed method CML+BDTR outperforms MACE.

5.5 Summary

We have developed a cross-domain face recognition method that focuses on thermal to visual image transfer using collaborative metric learning. Our approach incorporates maximum margin matrix factorization to learn a more generalised latent space that captures both intergroup and intragroup similarities effectively. By maximising the margin, we ensure that the learnt latent space is capable of generalising well. To establish the connection between image domains and the latent space, we employ deep architectures to derive the mapping function. This mapping function, combined with the learned latent space, forms the basis of our learned metric. Through alternate updation of the latent space and mapping function, we enhance the generalisation capability of our metric learning method.

To evaluate the performance of these methods, we conducted experiments on two

datasets: RGB-D-T and RegDB. The evaluation was performed in three learning settings: zero-shot, one-shot, and five-shot. In the RGB-D-T dataset, the test set comprised more than 7,000 image pairs from 25 individuals, while in the RegDB dataset, it included more than 1,000 image pairs. These methods demonstrated superior performance compared to state-of-the-art methods in the few-shot learning settings on both datasets.

CHAPTER 6

Conclusions & Future Work

In this thesis, our focus was on the development of novel techniques for thermal to visual cross-domain face recognition. We identified domain discrepancy and shortage of training data as the main challenges in cross-domain face recognition. To address these challenges, we proposed three methods: deep transfer learning, common subspace learning using dictionary learning, and collaborative metric learning.

The first method employed deep transfer learning, where the thermal classifier was trained by leveraging the knowledge from the visual classifier through transfer learning. We introduced sparsification of the network and transferred the weights of the sparsified network. Experimental results on the RGB-D-T dataset and UND-X1 collection demonstrated improved performance in thermal to visual face recognition, with an overall accuracy increased from 89.3% to 94.32% on the RGB-D-T dataset and from 81.54% to 90.33% on the UND-X1 dataset.

The second method involved a two-stage cross-domain face recognition approach based on dictionary learning. We projected the thermal and visual domain images onto a common subspace, where representation codes were used to describe the face images. In the second stage, metric learning was employed to measure the similarity between the representation codes. Experimental evaluations on the RGB-D-T and RegDB datasets demonstrated the effectiveness of the proposed method, with the CSL2+DML method outperforming others, even when there were no common individuals between the training and testing sets.

The third method introduced collaborative metric learning, where a latent space for

the metric was obtained using maximum margin matrix factorization, preserving the training similarities. The mapping from the image space to the learned latent space was achieved using a convolutional neural network. This method showed superior performance in the few-shot learning settings on both the RGB-D-T and RegDB datasets.

For future work, we aim to investigate effective sparsified network architectures that enhance generalization. Additionally, exploring kernel-based dictionary learning for obtaining a more generalized common subspace is an area of interest. We also intend to focus on advanced deep network architectures tailored for thermal to visual cross-domain face recognition, enhancing zero-shot learning capabilities, and incorporating self-attention and cross-attention mechanisms in cross-domain face recognition. By addressing these aspects, we can further advance the field and improve the accuracy and robustness of thermal to visual cross-domain face recognition systems.

List Of Papers Based On Thesis

- Gavini, Yaswanth, B. M. Mehtre, and Arun Agarwal. "Thermal to Visual Face Recognition using Transfer Learning." *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*. IEEE, 2019. https://doi.org/10.1109/ISBA.2019.8778474

- Gavini, Yaswanth, Arun Agarwal, and B. M. Mehtre. "Cross-Domain Face Recognition Using Dictionary Learning." *International Conference on Multi-disciplinary Trends in Artificial Intelligence*. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-33709-4_15

- Gavini, Yaswanth, Arun Agarwal, B. M. Mehtre. "Thermal to Visual Person Re-Identification Using Collaborative Metric Learning Based on Maximum Margin Matrix Factorization." *Pattern Recognition.* Volume 134, 2023, 109069, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2022.109069.

Papers submitted / to be communicated

- Gavini, Yaswanth, B. M. Mehtre, and Arun Agarwal. "Common-subspace learning for Cross-Domain Face Recognition Using Dictionary Learning."

- "Literature survey on thermal to visual face recognition "

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[3] Suzan Ece Ada, Emre Ugur, and H Levent Akin. Generalization in transfer learning. *arXiv preprint arXiv:1909.01331*, 2019.

[4] Jahanzeb Ahmad, Usman Ali, and Rashid Jalal Qureshi. Fusion of thermal and visual images for efficient face recognition using gabor filter. In *IEEE International Conference on Computer Systems and Applications, 2006.*, pages 135–139. IEEE Computer Society, 2006.

[5] Muna O. Almasawa, Lamiaa A. Elrefaei, and Kawthar Moria. A survey on deep learning-based person re-identification systems. *IEEE Access*, 7:175228–175247, 2019.

[6] Fernando Alonso-Fernandez, Kevin Hernandez-Diaz, Silvia Ramis, Francisco J Perales, and Josef Bigun. Facial masks and soft-biometrics: Leveraging face recognition cnns for age and gender prediction on mobile ocular images. *IET Biometrics*, 10(5):562–580, 2021.

[7] Slawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1571–1580, 2017.

[8] S. Bhattacharya and A. Routray. Heterogeneous face quality assessment. *Neural Computing and Applications*, 34(14):11589–11602, 2022.

[9] Mrinal Kanti Bhowmik, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu, and Mahantapas Kundu. Optimum fusion of visual and thermal face images for recognition. In *2010 Sixth International Conference on Information Assurance and Security*, pages 311–316. IEEE, 2010.

[10] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[11] Kevin W Bowyer, Kyong I Chang, Patrick J Flynn, and Xin Chen. Face recognition using 2-d, 3-d, and infrared: Is multimodal better than multisample? *Proceedings of the IEEE*, 94(11):2000–2012, 2006.

[12] Pradeep Buddharaju and Ioannis Pavlidis. Multispectral face recognition: fusion of visual imagery with physiological information. In *Face Biometrics for Personal Identification: Multi-Sensory Multi-Modal Systems*, pages 91–108. Springer, 2007.

[13] Sijia Cai, Wangmeng Zuo, Lei Zhang, Xiangchu Feng, and Ping Wang. Support vector guided dictionary learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 624–639, Cham, 2014. Springer International Publishing.

[14] Chi Ho Chan, Xuan Zou, Norman Poh, and Josef Kittler. Illumination invariant face recognition: a survey. In *Computer Vision: Concepts, Methodologies, Tools, and Applications*, pages 58–79. IGI Global, 2018.

[15] K.I. Chang, K.W. Bowyer, and P.J. Flynn. An evaluation of multimodal 2d+3d face biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):619–624, 2005.

[16] U. Cheema, M. Ahmad, D. Han, and S. Moon. Heterogeneous visible-thermal and visible-infrared face recognition using cross-modality discriminator network and unit-class loss. *Computational Intelligence and Neuroscience*, 2022, 2022.

[17] Cunjian Chen and Arun Ross. Matching thermal to visible face images using hidden factor analysis in a cascaded subspace learning framework. *Pattern Recognition Letters*, 72:25–32, 2016.

[18] Cunjian Chen and Arun Ross. Matching thermal to visible face images using a semantic-guided generative adversarial network. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[19] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[20] X. Chen, Patrick J. Flynn, and Kevin W. Bowyer. Visible-light and infrared face recognition. In *ACM Workshop on Multimodal User Authentication*, pages 48–55, 2003.

[21] Xuerong Chen, Zhongliang Jing, and Gang Xiao. Fuzzy fusion for face recognition. In *Fuzzy Systems and Knowledge Discovery: Second International Conference, FSKD 2005, Changsha, China, August 27-29, 2005, Proceedings, Part I 2*, pages 672–675. Springer, 2005.

[22] MyeongAh Cho, Taeoh Kim, Ig-Jae Kim, Kyungjae Lee, and Sangyoun Lee. Relational deep feature learning for heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 16:376–388, 2021.

[23] Jonghyun Choi, Shuowen Hu, S Susan Young, and Larry S Davis. Thermal to visible face recognition. In *Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring II; and Biometric Technology for Human Identification IX*, volume 8371, pages 252–261. Spie, 2012.

[24] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[25] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[26] Weihong Deng, Jun Guo, Jiani Hu, and Honggang Zhang. Comment on" 100% accuracy in automatic face recognition". *science*, 321(5891):912–912, 2008.

[27] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)*, 7(3):1–42, 2016.

[28] Sheng-Hung Fan, Min-Hong Lin, Jung-Yi Jiang, and Yau-Hwang Kuo. A few-shot learning method using feature reparameterization and dual-distance metric learning for object re-identification. *IEEE Access*, 9:133650–133662, 2021.

[29] Sajad Farokhi, Jan Flusser, and Usman Ullah Sheikh. Near infrared face recognition: A literature survey. *Computer Science Review*, 21:1–17, 2016.

[30] William A Firestone. Alternative arguments for generalizing from data as applied to qualitative research. *Educational researcher*, 22(4):16–23, 1993.

[31] Patrick J. Flynn, Kevin W. Bowyer, and P. Jonathon Phillips. Assessment of time dependency in face recognition: An initial study. In Josef Kittler and Mark S. Nixon, editors, *Audio- and Video-Based Biometric Person Authentication*, pages 44–51, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[32] Cedric Nimpa Fondje, Shuowen Hu, Nathaniel J. Short, and Benjamin S. Riggan. Cross-domain identification for thermal-to-visible face recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, Sep. 2020.

[33] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He. Dvg-face: Dual variational generation for heterogeneous face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2938–2952, 2022.

[34] C. Galea and R.A. Farrugia. Matching software-generated sketches to face photographs with a very deep cnn, morphed faces, and transfer learning. *IEEE Transactions on Information Forensics and Security*, 13(6):1421–1431, 2018.

[35] Cunyuan Gao, Rui Yao, Jiaqi Zhao, Yong Zhou, Fuyuan Hu, and Leida Li. Structure-aware person search with self-attention and online instance aggregation matching. *Neurocomputing*, 369:29–38, 2019.

[36] Yaswanth Gavini, Arun Agarwal, and B. M. Mehtre. Cross-domain face recognition using dictionary learning. In Rapeeporn Chamchong and Kok Wai Wong, editors, *Multi-disciplinary Trends in Artificial Intelligence*, pages 168–180, Cham, 2019. Springer International Publishing.

[37] Dihong Gong, Zhifeng Li, Weilin Huang, Xuelong Li, and Dacheng Tao. Heterogeneous face recognition: A common encoding feature discriminant approach. *Trans. Img. Proc.*, 26(5):2079–2089, May 2017.

[38] Guodong Guo, Lingyun Wen, and Shuicheng Yan. Face authentication with makeup changes. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):814–825, 2013.

[39] Guodong Guo and Na Zhang. A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805, 2019.

[40] Huimin Guo, Zhuolin Jiang, and Larry S. Davis. Discriminative dictionary learning with pairwise constraints. In Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, *Computer Vision – ACCV 2012*, pages 328–342, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[41] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[42] S. Haghiri, H. R. Rabiee, A. Soltani-Farani, S. A. Hosseini, and M. Shadloo. Locality preserving discriminative dictionary learning. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5242–5246, Oct 2014.

[43] Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Enhao Gong, Shijian Tang, Erich Elsen, Peter Vajda, Manohar Paluri, John Tran, Bryan Catanzaro, and William J. Dally. Dsd: Dense-sparse-dense training for deep neural networks. *International Conference on Learning Representations (ICLR)*, 2017.

[44] Yi Hao, Jie Li, Nannan Wang, and Xinbo Gao. Modality adversarial neural network for visible-thermal person re-identification. *Pattern Recognition*, 107:107533, 2020.

[45] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8385–8392, 2019.

[46] R. He, X. Wu, Z. Sun, and T. Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1761–1773, 2019. cited By 155.

[47] Ran He, Jie Cao, Lingxiao Song, Zhenan Sun, and Tieniu Tan. Adversarial cross-spectral face completion for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1025–1037, 2020.

[48] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Aasa Feragen, Marcello Pelillo, and Marco Loog, editors, *Similarity-Based Pattern Recognition*, pages 84–92, Cham, 2015. Springer International Publishing.

[49] Zheran Hong, Bin Liu, Yan Lu, Guojun Yin, and Nenghai Yu. Scale voting with pyramidal feature fusion network for person search. *IEEE Access*, 7:139692–139702, 2019.

[50] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1875–1882, 2014.

[51] Shuowen Hu, Jonghyun Choi, Alex L. Chan, and William Robson Schwartz. Thermal-to-visible face recognition using partial least squares. *J. Opt. Soc. Am. A*, 32(3):431–442, Mar 2015.

[52] Shuowen Hu, Prudhvi Gurram, Heesung Kwon, and Alex L Chan. Thermal-to-visible face recognition using multiple kernel learning. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, volume 9091, pages 356–362. SPIE, 2014.

[53] W. Hu and H. Hu. Disentangled spectrum variations networks for nir-vis face recognition. *IEEE Transactions on Multimedia*, 22(5):1234–1248, 2020.

[54] Weipeng Hu, Wenjun Yan, and Haifeng Hu. Dual face alignment learning network for nir-vis face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2411–2424, 2021.

[55] D. Huang and Y. F. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *2013 IEEE International Conference on Computer Vision*, pages 2496–2503, Dec 2013.

[56] M Indu and KV Kavitha. Survey on sketch based image retrieval methods. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pages 1–4. IEEE, 2016.

[57] Khawar Islam. Person search: New paradigm of person re-identification: A survey and outlook of recent works. *Image and Vision Computing*, 101:103970, 2020.

[58] Sina Jahanbin, Hyohoon Choi, and Alan C. Bovik. Passive multimodal 2-d+3-d face recognition using gabor features and landmark distances. *IEEE Transactions on Information Forensics and Security*, 6(4):1287–1304, 2011.

[59] Chaitra Jambigi, Ruchit Rawal, and Anirban Chakraborty. Mmd-reid: A simple but effective solution for visible-thermal person reid. In *British Machine Vision Conference*, 2021.

[60] Rob Jenkins and A Mike Burton. 100% accuracy in automatic face recognition. *Science*, 319(5862):435–435, 2008.

[61] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR 2011*, pages 1697–1704, June 2011.

[62] Dongoh Kang, Hu Han, Anil K Jain, and Seong-Whan Lee. Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching. *Pattern Recognition*, 47(12):3750–3766, 2014.

[63] Jin Kyu Kang, Toan Minh Hoang, and Kang Ryoung Park. Person re-identification between visible and thermal camera images based on deep residual cnn using single input. *IEEE Access*, 7:57972–57984, 2019.

[64] Alperen Kantarcı and Hazım Kemal Ekenel. Thermal to visible face recognition using deep autoencoders. In *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2019.

[65] Kavita Kavita and Rajender Singh Chhillar. Human face recognition and age estimation with machine learning: A critical review and future perspective. *International journal of electrical and computer engineering systems*, 13(10):945–952, 2022.

[66] Landry Kezebou, Victor Oludare, Karen Panetta, and Sos Agaian. Tr-gan: Thermal to rgb face synthesis with generative adversarial network for cross-modal face recognition. In *Mobile Multimedia/Image Processing, Security, and Applications 2020*, volume 11399, pages 158–168. SPIE, 2020.

[67] Sajid Ali Khan, Muhammad Ishtiaq, Muhammad Nazir, and Muhammad Shaheen. Face recognition under varying expressions and illumination using particle swarm optimization. *Journal of computational science*, 28:94–100, 2018.

[68] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[69] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[70] Shu Kong and Donghui Wang. A dictionary learning approach for classification: Separating the particularity and the commonality. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 186–199, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[71] Mate Krišto and Marina Ivasic-Kos. An overview of thermal face recognition methods. In *2018 41St international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1098–1103. IEEE, 2018.

[72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[73] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, June 2012.

[74] H.A. Le and I.A. Kakadiaris. Dblface: Domain-based labels for nir-vis heterogeneous face recognition. In *IJCB 2020 - IEEE/IAPR International Joint Conference on Biometrics*, 2020.

[75] Ha A Le and Ioannis A Kakadiaris. Illumination-invariant face recognition with deep relit face images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2146–2155. IEEE, 2019.

[76] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[77] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[78] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2020.

[79] F. Li and M. Jiang. Low-resolution face recognition and feature selection based on multidimensional scaling joint l2,1-norm regularisation. *IET Biometrics*, 8(3):198–205, 2019.

[80] Wei-Hong Li, Yafang Mao, Ancong Wu, and Wei-Shi Zheng. Correlation based identity filter: An efficient framework for person search. In Yao Zhao, Xiangwei Kong, and David Taubman, editors, *Image and Graphics*, pages 250–261, Cham, 2017. Springer International Publishing.

[81] W. Liang, G. Wang, J. Lai, and X. Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE Transactions on Image Processing*, 30:6392–6407, 2021.

[82] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, and Nicu Sebe. Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In *Proceedings of the 28th ACM international conference on multimedia*, pages 889–897, 2020.

[83] Haijun Liu, Jian Cheng, Wen Wang, Yanzhou Su, and Haiwei Bai. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing*, 398:11–19, 2020.

[84] Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 23:4414–4425, 2020.

[85] Weiyang Liu, Zhiding Yu, Lijia Lu, Yandong Wen, Hui Li, and Yuexian Zou. Kcrc-lcd: Discriminative kernel collaborative representation with locality constrained dictionary for visual categorization. *Pattern Recognition*, 48(10):3076 – 3092, 2015. Discriminative Feature Learning from Big Data for Visual Recognition.

[86] Xin Liu, Shiguang Shan, and Xilin Chen. Face recognition after plastic surgery: A comprehensive study. In Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, *Computer Vision – ACCV 2012*, pages 565–576, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[87] Angelique Loesch, Jaonary Rabarisoa, and Romaric Audigier. End-to-end person search sequentially trained on aggregated dataset. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4574–4578, 2019.

[88] F. Ma, X.-Y. Jing, L. Cheng, S. Wu, H. Zhang, Y. Yao, and X. Zhu. Person re-identification with character-illustration-style image and normal photo. *IEEE Access*, 9:30486–30495, 2021.

[89] Zahid Mahmood, Tauseef Ali, and Samee U Khan. Effects of pose and image resolution on automatic face recognition. *IET biometrics*, 5(2):111–119, 2016.

[90] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, March 2010.

[91] Zuheng Ming, Junshi Xia, Muhammad Muzzamil Luqman, Jean-Christophe Burie, and Kaixing Zhao. Dynamic multi-task learning for face recognition with facial expression. In *Lightweight Face Recognition Challenge Workshop during the 2019 International Conference on Computer Vision (ICCV 2019)*, 2019.

[92] T. Miyamoto, H. Hashimoto, A. Hayasaka, A.F. Ebihara, and H. Imaoka. Joint feature distribution alignment learning for nir-vis and vis-vis face recognition. In *2021 IEEE International Joint Conference on Biometrics, IJCB 2021*, 2021.

[93] S. P. Mudunuri and S. Biswas. Dictionary alignment for low-resolution and heterogeneous face recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1115–1123, March 2017.

[94] Sivaram Prasad Mudunuri and Soma Biswas. Coarse to fine training for low-resolution heterogeneous face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2421–2425, 2018.

[95] Sivaram Prasad Mudunuri, Shashanka Venkataramanan, and Soma Biswas. Improved low resolution heterogeneous face recognition using re-ranking. In *Computer Vision, Pattern Recognition, Image Processing, and Graphics: 6th National Conference, NCVPRIPG 2017, Mandi, India, December 16-19, 2017, Revised Selected Papers 6*, pages 446–456. Springer, 2018.

[96] S. Nagpal, M. Singh, R. Singh, and M. Vatsa. Discriminative shared transform learning for sketch to image matching. *Pattern Recognition*, 114, 2021.

[97] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[98] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[99] Olegs Nikisins, Kamal Nasrollahi, Modris Greitans, and Thomas B Moeslund. Rgb-dt based face recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1716–1721. IEEE, 2014.

[100] Xin Ning, Fangzhe Nan, Shaohui Xu, Lina Yu, and Liping Zhang. Multi-view frontal face image generation: a survey. *Concurrency and Computation: Practice and Experience*, page e6147, 2020.

[101] Beom-Seok Oh, Kangrok Oh, Andrew Beng Jin Teoh, Zhiping Lin, and Kar-Ann Toh. A gabor-based network for heterogeneous face recognition. *Neurocomputing*, 261:253 – 265, 2017. Advances in Extreme Learning Machines (ELM 2015).

[102] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, June 2014.

[103] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, Xueming Li, Chen Change Loy, and Xiaogang Wang. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. *Image and Vision Computing*, 56:28 – 48, 2016.

[104] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[105] M. Pang, B. Wang, S. Chen, Y.-M. Cheung, R. Zou, and W. Huang. Cross-domain prototype learning from contaminated faces via disentangling latent factors. In *International Conference on Information and Knowledge Management, Proceedings*, pages 4369–4373, 2022.

[106] M. Pang, B. Wang, S. Huang, Y.-M. Cheung, and B. Wen. A unified framework for bidirectional prototype learning from contaminated faces across heterogeneous domains. *IEEE Transactions on Information Forensics and Security*, 17:1544–1557, 2022.

[107] Hemprasad Y Patil, Ashwin G Kothari, and Kishor M Bhurchandi. Expression invariant face recognition using local binary patterns and contourlet transform. *Optik*, 127(5):2670–2678, 2016.

[108] Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, Duc Thanh Nguyen, Gemma Roig, and Sai-Kit Yeung. Lcd: Learned cross-domain descriptors for 2d-3d matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11856–11864, 2020.

[109] Chikontwe Philip and Hyo Jong Lee. Face sketch synthesis : A neural style approach. In *Int'l Conf on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, 2017.

[110] Denise F Polit and Cheryl Tatano Beck. Generalization in quantitative and qualitative research: Myths and strategies. *International journal of nursing studies*, 47(11):1451–1458, 2010.

[111] Q. Qiu and R. Chellappa. Compositional dictionaries for domain adaptive face recognition. *IEEE Transactions on Image Processing*, 24(12):5152–5165, Dec 2015.

[112] K Rajesh and Atul Negi. Heuristic based learning of parameters for dictionaries in sparse representations. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1013–1019. IEEE, 2018.

[113] C. Reale, H. Lee, and H. Kwon. Deep heterogeneous face recognition networks based on cross-modal distillation and an equitable distance metric. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 226–232, July 2017.

[114] C. Reale, N. M. Nasrabadi, and R. Chellappa. Coupled dictionaries for thermal to visible face recognition. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 328–332, Oct 2014.

[115] Samira Reihanian, Ehsan Arbabi, and Behrouz Maham. Random sparse representation for thermal to visible face recognition. In *2017 IEEE Symposium on Computers and Communications (ISCC)*, pages 1380–1385. IEEE, 2017.

[116] S. Reyhanian and E. Arbabi. Weighted vote fusion in prototype random subspace for thermal to visible face recognition. In *2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 1–5, March 2015.

[117] B. S. Riggan, C. Reale, and N. M. Nasrabadi. Coupled auto-associative neural networks for heterogeneous face recognition. *IEEE Access*, 3:1620–1632, 2015.

[118] Lorenzo Rosasco, Alessandro Verri, Matteo Santoro, Sofia Mosci, and Silvia Villa. Iterative projection methods for structured sparsity regularization. 2009.

[119] H. Roy and D. Bhattacharjee. A novel local wavelet energy mesh pattern (lwemep) for heterogeneous face recognition. *Image and Vision Computing*, 72:1–13, 2018.

[120] M. Saquib Sarfraz and Rainer Stiefelhagen. Deep Perceptual Mapping for Cross-Modal Face Recognition. *International Journal of Computer Vision*, 122(3):426–438, 2017.

[121] Manisha M Sawant and Kishor M Bhurchandi. Age invariant face recognition: a survey on facial aging databases, techniques and effect of aging. *Artificial Intelligence Review*, 52:981–1008, 2019.

[122] Wei Shi, Hong Liu, Fanyang Meng, and Weipeng Huang. Instance enhancing loss: Deep identity-sensitive feature embedding for person search. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4108–4112, 2018.

[123] Zhiyuan Shi, Timothy M. Hospedales, and Tao Xiang. Transferring a semantic representation for person re-identification and search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[124] M. O. Simón, C. Corneanu, K. Nasrollahi, O. Nikisins, S. Escalera, Y. Sun, H. Li, Z. Sun, T. B. Moeslund, and M. Greitans. Improved rgb-d-t based face recognition. *IET Biometrics*, 5(4):297–303, 2016.

[125] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. Derivenet for (very) low resolution image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6569–6577, 2022.

[126] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. Disguise resilient face verification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3895–3905, 2022.

[127] Maneet Singh, Shruti Nagpal, Richa Singh, and Mayank Vatsa. Dual directed capsule network for very low resolution image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 340–349, 2019.

[128] Richa Singh, Mayank Vatsa, Himanshu S. Bhatt, Samarth Bharadwaj, Afzel Noore, and Shahin S. Nooreyezdan. Plastic surgery: A new dimension to face recognition. *IEEE Transactions on Information Forensics and Security*, 5(3):441–448, 2010.

[129] Xulin Song and Zhong Jin. Domain adaptive attention-based dropout for one-shot person re-identification. *International Journal of Machine Learning and Cybernetics*, 13(1):255–268, 2022.

[130] Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005.

[131] Lin Sun and Zengwei Zheng. Thermal-to-visible face alignment on edge map. *IEEE Access*, 5:11215–11227, 2017.

[132] Z. Sun, C. Fu, M. Luo, and R. He. Self-augmented heterogeneous face recognition. In *2021 IEEE International Joint Conference on Biometrics, IJCB 2021*, 2021.

[133] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[134] T. H. Vu and V. Monga. Learning a low-rank shared dictionary for object classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4428–4432, Sep. 2016.

[135] T. H. Vu and V. Monga. Fast low-rank shared dictionary learning for image classification. *IEEE Transactions on Image Processing*, 26(11):5160–5175, Nov 2017.

[136] Donghui Wang and Shu Kong. A classification-oriented dictionary learning model: Explicitly learning the particularity and commonality across categories. *Pattern Recognition*, 47(2):885 – 898, 2014.

[137] Hao Wang, Dihong Gong, Zhifeng Li, and Wei Liu. Decorrelated adversarial learning for age-invariant face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3527–3536, 2019.

[138] Jiabao Wang, ShanShan Jiao, Yang Li, and Zhuang Miao. Two-stage metric learning for cross-modality person re-identification. In *Proceedings of the 5th International Conference on Multimedia and Image Processing*, pages 28–32, 2020.

[139] Nannan Wang, Xinbo Gao, Leiyu Sun, and Jie Li. Bayesian face sketch synthesis. *IEEE Transactions on Image Processing*, 26(3):1264–1274, 2017.

[140] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *International Journal of Computer Vision*, 106(1):9–30, Jan 2014.

[141] Pingyu Wang, Fei Su, Zhicheng Zhao, Yanyun Zhao, Lei Yang, and Yang Li. Deep hard modality alignment for visible thermal person re-identification. *Pattern Recognition Letters*, 133:195–201, 2020.

[142] Pingyu Wang, Zhicheng Zhao, Fei Su, Yanyun Zhao, Haiying Wang, Lei Yang, and Yang Li. Deep multi-patch matching network for visible thermal person re-identification. *IEEE Transactions on Multimedia*, 23:1474–1488, 2020.

[143] Xiaolong Wang and Chandra Kambhamettu. A new approach for face recognition under makeup changes. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 423–427. IEEE, 2015.

[144] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. Language person search with mutually connected classification loss. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2057–2061, 2019.

[145] Zhongling Wang, Zhenzhong Chen, and Feng Wu. Thermal to visible facial image translation using generative adversarial networks. *IEEE Signal Processing Letters*, 25(8):1161–1165, 2018.

[146] C. Wei and Y. F. Wang. Undersampled face recognition via robust auxiliary dictionary learning. *IEEE Transactions on Image Processing*, 24(6):1722–1734, June 2015.

[147] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016.

[148] Xiang Wu, Lingxiao Song, Ran He, and Tieniu Tan. Coupled deep learning for heterogeneous face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[149] Yong Wu, Sizhe Wan, Di Wu, Chao Wang, Changan Yuan, Xiao Qin, Hongjie Wu, and Xingming Zhao. Position attention-guided learning for infrared-visible person re-identification. In *Intelligent Computing Theories and Application: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part I 16*, pages 387–397. Springer, 2020.

[150] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.

[151] X. Xu, Y. Li, and Y. Jin. Hierarchical discriminant feature learning for cross-modal face recognition. *Multimedia Tools and Applications*, 79(45-46):33483–33502, 2020.

[152] Xiaolin Xu, Yidong Li, Yi Jin, Congyan Lang, Songhe Feng, and Tao Wang. Hierarchical discriminant feature learning for heterogeneous face recognition. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2018.

[153] Y. Xu, Z. Li, J. Yang, and D. Zhang. A survey of dictionary learning algorithms for face recognition. *IEEE Access*, 5:8502–8514, 2017.

[154] Meng Yang, Weiyang Liu, Weixin Luo, and Linlin Shen. Analysis-synthesis dictionary learning for universality-particularity representation based classification. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[155] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Sparse representation based fisher discrimination dictionary learning for image classification. *International Journal of Computer Vision*, 109(3):209–232, Sep 2014.

[156] Z. Yang, J. Liang, C. Fu, M. Luo, and X.-Y. Zhang. Heterogeneous face recognition via face synthesis with identity-attribute disentanglement. *IEEE Transactions on Information Forensics and Security*, 17:1344–1358, 2022.

[157] Mang Ye, Xiangyuan Lan, and Qingming Leng. Modality-aware collaborative learning for visible thermal person re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 347–355, 2019.

[158] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29:9387–9399, 2020.

[159] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[160] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019.

[161] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 229–247, Cham, 2020. Springer International Publishing.

[162] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, page 2, 2018.

[163] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3320–3328, Cambridge, MA, USA, 2014. MIT Press.

[164] Sulan Zhai, Shunqiang Liu, Xiao Wang, and Jin Tang. Fmt: fusing multi-task convolutional neural network for person search. *Multimedia Tools and Applications*, 78(22):31605–31616, 2019.

[165] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2021.

[166] T. Zhang, A. Wiliem, S. Yang, and B. Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *2018 International Conference on Biometrics (ICB)*, pages 174–181, Feb 2018.

[167] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022.

[168] D. Zhao, Z. Chen, C. Liu, and Y. Peng. Two-dimensional linear discriminant analysis for low-resolution face recognition. In *Proceedings - 2017 Chinese Automation Congress, CAC 2017*, volume 2017-January, pages 703–707, 2017.

[169] Feng Zhao, Jing Li, Lu Zhang, Zhe Li, and Sang-Gyun Na. Multi-view face recognition using deep neural networks. *Future Generation Computer Systems*, 111:375–380, 2020.

[170] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018.

[171] Sanqiang Zhao, Yongsheng Gao, and Baochang Zhang. Gabor feature constrained statistical model for efficient landmark localization and face recognition. *Pattern Recognition Letters*, 30(10):922–930, 2009.

[172] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 336–341, 1998.

[173] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[174] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):591–606, 2016.

[175] Y. Zheng, D. Huang, W. Li, S. Wang, and Y. Wang. 2d-3d heterogeneous face recognition based on deep coupled spectral regression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2019-June, pages 277–284, 2019.

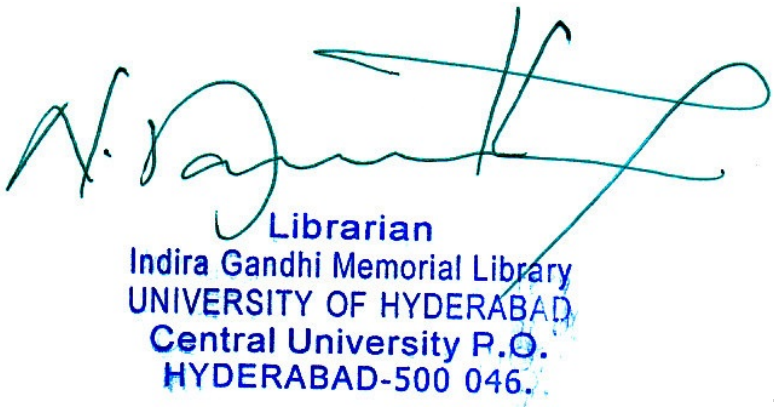
[176] H. Zhou, A. Mian, L. Wei, D. Creighton, M. Hossny, and S. Nahavandi. Recent advances on singlemodal and multimodal face recognition: A survey. *IEEE Transactions on Human-Machine Systems*, 44(6):701–716, Dec 2014.

[177] Fan Zhu and Ling Shao. Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision*, 109(1):42–59, Aug 2014.

[178] Jun-Yong Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Logarithm gradient histogram: A general illumination invariant descriptor for face recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

Thermal to Visual: A Cross-Domain Face Recognition System

by Yaswanth Gavini



Submission date: 30-Jun-2023 10:35AM (UTC+0530)
Submission ID: 2124630836
File name: Yaswanth_Gavini.pdf (11.68M)
Word count: 27497
Character count: 146347

Thermal to Visual: A Cross-Domain Face Recognition System

ORIGINALITY REPORT

36%	13%	34%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

Net Similarity Index = 09

Noted
Attested
30/6/23
DEAN
School of CIS
Dr. C R. Rao Road,
Central University Campus PO
Gachibowli Hyderabad-46. (India)

1	Yaswanth Gavini, Arun Agarwal, B.M. Mehtre. "Thermal to Visual Person Re-Identification Using Collaborative Metric Learning Based on Maximum Margin Matrix Factorization", Pattern Recognition, 2023 Publication	10%
2	Yaswanth Gavini, B. M. Mehtre, Arun Agarwal. "Thermal to Visual Face Recognition using Transfer Learning", 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA), 2019 Publication	9%
3	dokumen.pub Internet Source	6%
4	ipfs.io Internet Source	1%
5	"Multi-disciplinary Trends in Artificial Intelligence", Springer Science and Business Media LLC, 2019 Publication	1%

Sl.nos 1 to 5 are connected to our papers published.

www.researchgate.net

In the similarity report, the following five sources are from the student's own publications. The details are given below:

1. **Source-1:** 10% of the similarity is from the student's following publication.
    a. Gavini, Yaswanth, Arun Agarwal, and B. M. Mehtre. "Thermal to Visual Person Re-Identification Using Collaborative Metric Learning Based on Maximum Margin Matrix Factorization." *Pattern Recognition* 134 (2023): 109069, doi:10.1016/j.patcog.2022.109069.

2. **Source- 2:** 9% of the similarity is from the student's below publication.
    a. Gavini, Yaswanth, B. M. Mehtre, and Arun Agarwal. "Thermal to visual face recognition using transfer learning." In *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pp. 1-8. IEEE, 2019, doi: 10.1109/ISBA.2019.8778474.

3. **Source-3:** 6% of the similarity is from the internet source "dokumen.pub", and the "dokumen.pub" portal has the student's publication at the following link: https://dokumen.pub/multi-disciplinary-trends-in-artificial-intelligence-13th-international-conference-miwai-2019-kuala-lumpur-malaysia-november-1719-2019-proceedings-1st-ed-2019-978-3-030-33708-7-978-3-030-33709-4.html
4. **Source-4:** 1% of the similarity comes from the internet source "ipfs.io". The "ipfs.io" portal has the student's publication at the following link: https://ipfs.io/ipfs/bafykbzacebvlhbhe5rgolrb35puigvn24jxivlpvxo34brq2zzcayubc3q3fi?filename=%5BLecture+Notes+in+Computer+Science+%E2%84%9611909%5D+Rapeeporn+Chamchong%2C+Kok+Wai+Wong+-+Multi-disciplinary+Trends+in+Artificial+Intelligence%3A+13th+International+Conference%2C+MIWAI+2019%2C+Kuala+Lumpur%2C+Malaysia%2C+November+17%E2%80%9319%2C+2019%2C+Proceedings+%282019%2C+Springer%29+%5B10.1007%2F978-3-030-33709-4%5D.pdf
5. **Source-5:** 1% of the similarity is from the student's publication.

    Here, Source-3, Source-4, and Source-5 are sources in which the similarity is from the student's below publication.
    a. Gavini, Yaswanth, Arun Agarwal, and Babu M. Mehtre. "Cross-domain face recognition using dictionary learning." In *Multi-disciplinary Trends in Artificial Intelligence: 13th International Conference, MIWAI 2019, Kuala Lumpur, Malaysia, November 17–19, 2019, Proceedings 13*, pp. 168-180. Springer International Publishing, 2019, doi:10.1007/978-3-030-33709-4_15.

Noted

30/6/23

DEAN
School of CIS
Dr. C R. Rao Road,
Central University Campus PO
Gachibowli Hyderabad-46. (India)

6	Internet Source	1 %
7	Yong Xu, Zhengming Li, Jian Yang, David Zhang. "A Survey of Dictionary Learning Algorithms for Face Recognition", IEEE Access, 2017 Publication	<1 %
8	Lecture Notes in Computer Science, 2015. Publication	<1 %
9	Dorra Mahouachi, Moulay A. Akhloufi. "Recent Advances in Infrared Face Analysis and Recognition with Deep Learning", AI, 2023 Publication	<1 %
10	ijarcsse.com Internet Source	<1 %
11	"13th International Conference on Theory and Application of Fuzzy Systems and Soft Computing — ICAFS-2018", Springer Science and Business Media LLC, 2019 Publication	<1 %
12	Lecture Notes in Computer Science, 2013. Publication	<1 %
13	orca.cf.ac.uk Internet Source	<1 %
14	"Advances in Artificial Intelligence", Springer Science and Business Media LLC, 2019	<1 %

	Publication	
15	Submitted to Fachhochschule Kärnten Gemeinnützige Privatstiftung Student Paper	<1 %
16	repository.iiitd.edu.in Internet Source	<1 %
17	Siavash Haghiri, Hamid R. Rabiee, Ali Soltani-Farani, Seyyed Abbas Hosseini, Maryam Shadloo. "Locality preserving discriminative dictionary learning", 2014 IEEE International Conference on Image Processing (ICIP), 2014 Publication	<1 %
18	Submitted to University of Adelaide Student Paper	<1 %
19	"Advances in Brain Inspired Cognitive Systems", Springer Science and Business Media LLC, 2018 Publication	<1 %
20	deepai.org Internet Source	<1 %
21	www.mdpi.com Internet Source	<1 %
22	hcis-journal.springeropen.com Internet Source	<1 %
23	Zan-Xia Jin, Bo-Wen Zhang, Fang Zhou, Jingyan Qin, Xu-Cheng Yin. "Ranking via partial	<1 %

	ordering for answer selection", Information Sciences, 2020 Publication	
24	vdoc.pub Internet Source	<1 %
25	"Biometric Recognition", Springer Nature, 2016 Publication	<1 %
26	www.cse.cuhk.edu.hk Internet Source	<1 %
27	Submitted to Cranfield University Student Paper	<1 %
28	Submitted to University of Edinburgh Student Paper	<1 %
29	Giuseppe Schirripa Spagnolo. "Banknote security using a biometric-like technique: a hylemetric approach", Measurement Science and Technology, 05/01/2010 Publication	<1 %
30	Lecture Notes in Computer Science, 2012. Publication	<1 %
31	escholarship.org Internet Source	<1 %
32	tmukul.com Internet Source	<1 %

33	"Computer Vision – ACCV 2018", Springer Science and Business Media LLC, 2019 Publication	<1 %
34	Bo Geng, , Dacheng Tao, and Chao Xu. "DAML: Domain Adaptation Metric Learning", IEEE Transactions on Image Processing, 2011. Publication	<1 %
35	cerne.ufla.br Internet Source	<1 %
36	Fei Li, Mingyan Jiang. "Low-resolution face recognition and feature selection based on multidimensional scaling joint L 2,1-norm regularisation", IET Biometrics, 2019 Publication	<1 %
37	Chandrasekar Vuppalapati. "Machine Learning and Artificial Intelligence for Agricultural Economics", Springer Science and Business Media LLC, 2021 Publication	<1 %
38	Surya Prakash, Pei Yean Lee, Antonio Robles-Kelly. "Stereo techniques for 3D mapping of object surface temperatures", Quantitative InfraRed Thermography Journal, 2007 Publication	<1 %
39	Zhang, Dongyu, Pengju Liu, Kai Zhang, Hongzhi Zhang, Qing Wang, and Xiaoyuan Jing. "Class relatedness oriented-	<1 %

discriminative dictionary learning for multiclass image classification", Pattern Recognition, 2015.
Publication

40	hufee.meraka.org.za Internet Source	<1%
41	"Pattern Recognition and Computer Vision", Springer Science and Business Media LLC, 2019 Publication	<1%
42	"Data Management Technologies and Applications", Springer Science and Business Media LLC, 2018 Publication	<1%
43	"Meta-learning basics and background", Elsevier BV, 2023 Publication	<1%
44	Submitted to Associatie K.U.Leuven Student Paper	<1%
45	Jean-Paul Ainam, Ke Qin, Guisong Liu, Guangchun Luo. "View-Invariant and Similarity Learning for Robust Person Re-Identification", IEEE Access, 2019 Publication	<1%
46	www.arxiv-vanity.com Internet Source	<1%

47	"Computer Analysis of Images and Patterns", Springer Science and Business Media LLC, 2019 Publication	<1 %
48	Yaswanth Gavini, Arun Agarwal, B.M. Mehtre. "Thermal to Visual Person Re-Identification Using Collaborative Metric Learning Based on Maximum Margin Matrix Factorization", Pattern Recognition, 2022 Publication	<1 %
49	opus.lib.uts.edu.au Internet Source	<1 %
50	www.inventiva.co.in Internet Source	<1 %
51	"Image and Graphics Technologies and Applications", Springer Science and Business Media LLC, 2020 Publication	<1 %
52	Guodong Ding, Salman Khan, Zhenmin Tang, Fatih Porikli. "Feature mask network for person re-identification", Pattern Recognition Letters, 2019 Publication	<1 %
53	Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W. Tsang. "A deep learning framework for Hybrid Heterogeneous Transfer Learning", Artificial Intelligence, 2019	<1 %

Publication

54	Pingyu Wang, Fei Su, Zhicheng Zhao, Yanyun Zhao, Lei Yang, Yang Li. "Deep hard modality alignment for visible thermal person re-identification", Pattern Recognition Letters, 2020 Publication	<1 %
55	Si Si, Dacheng Tao, Kwok-Ping Chan. "Evolutionary Cross-Domain Discriminative Hessian Eigenmaps", IEEE Transactions on Image Processing, 2010 Publication	<1 %
56	Submitted to University of Hyderabad, Hyderabad Student Paper	<1 %
57	fr.scribd.com Internet Source	<1 %
58	Iman Deznabi, Busra Arabaci, Mehmet Koyutürk, Oznur Tastan. "DeepKinZero: Zero-Shot Learning for Predicting Kinase-Phosphosite Associations Involving Understudied Kinases", Cold Spring Harbor Laboratory, 2019 Publication	<1 %
59	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1 %

60	Submitted to University of Ulster Student Paper	<1 %
61	Zhengming Li, Zheng Zhang, Zizhu Fan, Jie Wen. "An interactively constrained discriminative dictionary learning algorithm for image classification", Engineering Applications of Artificial Intelligence, 2018 Publication	<1 %
62	www.semanticscholar.org Internet Source	<1 %
63	"Image and Graphics", Springer Science and Business Media LLC, 2017 Publication	<1 %
64	"International Conference on Advanced Intelligent Systems for Sustainable Development", Springer Science and Business Media LLC, 2023 Publication	<1 %
65	"Pattern Recognition and Computer Vision", Springer Science and Business Media LLC, 2018 Publication	<1 %
66	Communications in Computer and Information Science, 2015. Publication	<1 %
67	E.J.C. Kelkboom. "Binary Biometrics: An Analytic Framework to Estimate the Bit Error	<1 %

Probability under Gaussian Assumption", 2008 IEEE Second International Conference on Biometrics Theory Applications and Systems, 09/2008
Publication

68	Encyclopedia of Biometrics, 2015. Publication	<1 %
69	Submitted to SASTRA University Student Paper	<1 %
70	eprints.nottingham.ac.uk Internet Source	<1 %
71	"Intelligent Computing Theories and Application", Springer Science and Business Media LLC, 2021 Publication	<1 %
72	Submitted to Florida Atlantic University Student Paper	<1 %
73	Mang Ye, Xiangyuan Lan, Zheng Wang, Pong C. Yuen. "Bi-Directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification", IEEE Transactions on Information Forensics and Security, 2020 Publication	<1 %
74	Submitted to University of Westminster Student Paper	<1 %
75	jos.org.cn Internet Source	

		<1%
76	"Advances in Multimedia Information Processing - PCM 2016", Springer Science and Business Media LLC, 2016 Publication	<1%
77	"Artificial Neural Networks and Machine Learning – ICANN 2019: Deep Learning", Springer Science and Business Media LLC, 2019 Publication	<1%
78	"Computer Vision - ACCV 2014 Workshops", Springer Nature, 2015 Publication	<1%
79	Bo Li, Xiaohong Wu, Qiang Liu, Xiaohai He, Fei Yang. "Visible Infrared Cross-Modality Person Re-Identification Network Based on Adaptive Pedestrian Alignment", IEEE Access, 2019 Publication	<1%
80	Submitted to Canberra Institute of Technology Student Paper	<1%
81	Data-Driven Optimization and Knowledge Discovery for an Enterprise Information System, 2015. Publication	<1%

82	Junlin Hu, Jiwen Lu, Yap-Peng Tan, Jie Zhou. "Deep Transfer Metric Learning", IEEE Transactions on Image Processing, 2016 Publication	<1 %
83	link.springer.com Internet Source	<1 %
84	plaindata.blogspot.com Internet Source	<1 %
85	www.v7labs.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		