2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Cross-spectrum Face Recognition Using Subspace Projection Hashing

Hanrui Wang, Xingbo Dong, Zhe Jin
School of Information Technology
Monash University Malaysia
47500 Subang Jaya, Selangor, Malaysia
{hanrui.wang, xingbo.dong, jin.zhe}@monash.edu

Jean-Luc Dugelay
Department of Digital Security
EURECOM
Sophia-Antipolis, France
jld@eurecom.fr

Massimo Tistarelli
The University of Sassari
Alghero, SS 07041, Italy
tista@uniss.it

***Abstract*—Cross-spectrum face recognition, e.g. visible to thermal matching, remains a challenging task due to the large variation originated from different domains. This paper proposed a subspace projection hashing (SPH) to enable the cross-spectrum face recognition task. The intrinsic idea behind SPH is to project the features from different domains onto a *common subspace*, where matching the faces from different domains can be accomplished. Notably, we proposed a new loss function that can (i) preserve both inter-domain and intra-domain similarity; (ii) regularize a scaled-up pairwise distance between hashed codes, to optimize projection matrix. Three datasets, Wiki, EURECOM VIS-TH paired face and TDFace are adopted to evaluate the proposed SPH. The experimental results indicate that the proposed SPH outperforms the original linear subspace ranking hashing (LSRH) in the benchmark dataset (Wiki) and demonstrates a reasonably good performance for visible-thermal, visible-near-infrared face recognition, therefore suggests the feasibility and effectiveness of the proposed SPH.**

***Index Terms*—Cross-spectrum face recognition, visible to thermal, visible to near-infrared, subspace projection hashing**

I. Introduction

Face recognition (FR) operated on visible light has gained a long-standing interest attributed to the comfort and non-intrusive face image acquisition. Furthermore, recent advances in deep neural network achieved the superior accuracy of FR, e.g. Google FaceNet achieved 99.63% recognition accuracy on Labelled Faces in the Wild (LFW) dataset [1]. On the other hand, face images are often captured in other spectrum bands such as thermal or near-infrared. Cross-spectrum FR performs the matching of the face images from different modalities or domains, e.g. visible to thermal, visible to near-infrared. In this paper, the terms of "spectrum", "domain" and "modality" are used interchangeably. Considering the scenario that thermal face image could be effectively matched against the visible face image of the same identity, it will be a great interest for various applications such as surveillance [2] and forensics [3].

However, the performance of cross-spectrum FR can be largely varied due to the variety of modalities. For instance, visible (VIS) and thermal (TH) face demonstrate a large modality variation. The accuracy of cross-spectrum FR between visible (VIS) and thermal (TH) images therefore can be as lower as 60% [4], [5]. Whilst, the recognition rate of FR between visible and near-infrared (NIR) images can achieve 99.39% accuracy [6]. It can be seen that the matching performance of VIS-NIR is significantly accurate compared to the matching performance of VIS-TH. Zhang et al. [7] justified this large discrepancy of performance that thermal image lost the most of texture and edge information for the face images, which leads to the poor accuracy, while near-infrared image preserves the most similarities of visible image. This indicates that thermal-visible FR is more challenging than NIR-visible FR.

An ideal cross-spectrum FR aims several criteria: (i) a cross-spectrum FR system should achieve good accuracy under inter-domain situation; (ii) the matching accuracy under intra-domain should be preserved simultaneously; (iii) the matching process should be accomplished efficiently under the larger scale searching scenario.

Generative adversarial networks (GAN) based approach is one of the most popular approach for cross-spectrum FR. GAN synthesizes the thermal or near-infrared face images based on the corresponding visible face images, while it's hard to generate high quality thermal images by GAN with inadequate training samples; thus cannot guarantee the matching performance [5]. Also, training of GAN is computationally expensive and hard to converge [7], [8].

On the other hand, common subspace projection (CSP) (also known as Hashing) is another well-known approach for cross-spectrum FR. CSP projects features extracted from different domains onto a common subspace [9], [10]. Matching from different modalities can be performed in an identical subspace. CSP requires neither network training nor large datasets. Therefore, hashing may be a decent option that can be adopted for cross-spectrum FR problem. However, the existing hashing solutions for inter-domain retrieval such as canonical correlational analysis (CCA) [11], bilinear model (BLM) [12] and partial least squares (PLS) [13] merely consider inter-domain similarity measures on common subspace. An ideal inter-domain hashing requires that the samples from the same identity are as close as possible while those from different identities are far separated in the projected subspace [10]. These methods may perform cross-spectrum FR suboptimal when intra-class variation is large. Therefore, both inter-domain and intra-domain matching should be taken into account for cross-spectrum FR.

Inspired from the hashing method for media retrieval, i.e.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

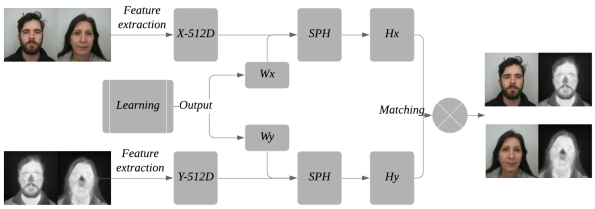


Fig. 1. The proposed SPH framework.

linear subspace ranking hashing (LSRH) [14], we proposed a new hashing method namely subspace projection hashing (SPH). A block diagram of SPH is illustrated in Fig. 1: (i) face features are extracted as 512-dimensional vectors; (ii) the hashing matrix W_X and W_Y is generated by learning process; (iii) the features are projected onto common subspace with hashing matrix, to generate hashed codes H_X and H_Y; (iv) the matching process is executed with hashed codes. SPH projects the face features extracted from different domains, i.e. visible, thermal and near-infrared onto a common subspace, thus realizes cross-spectrum FR with reasonable gain of accuracy. The projection matrix is learned based on a loss function for inter-domain and intra-domain matching as discussed in section IV. The generated hashed code is in binary/integer form which is fast for matching [15], efficient for data storage and retrieval [16] due to the pure involvement of the bit-wise operations.

The contributions of this paper are:

- We proposed a new hashing method for cross-spectrum FR task, namely SPH. The proposed SPH achieved satisfied accuracy of FR on visible-thermal and visible-NIR spectrum.
- A new loss function that optimizes both the inter-domain and intra-domain similarities are introduced. Specifically, visible-thermal and/or visible-near-infrared, and visible-visible and/or thermal-thermal. The new loss illustrates an improved performance with respect to the original LSRH.
- We performed a series of experiments on visible-thermal, visible-NIR for FR. The proposed method shows a reasonably good performance and validate the effectiveness of the proposed method for cross-spectrum FR.

II. Related Work

In literature, a number of methods were proposed to attempt cross-spectrum FR challenge. These methods can be generally divided into three distinct approaches [17], i.e. (i) hand-crafted based approach, (ii) generative adversarial network (GAN) based approach, and (iii) subspace projection (also known as hashing) based approach. In this paper, we focus on the subspace projection based approach. Hence, the related works are restricted to the subspace projection based approach. Briefly, the basic idea of subspace projection approach is to seek a common subspace for multi-domain feature/variables, and project each feature/variable from different domain onto the common subspace. The projection can be completed by two different mechanisms, (i) identified correlations, such as canonical correlation analysis (CCA) [11] and partial least squares (PLS) [13]), (ii) similarity labels, e.g. generalized multi-view analysis (GMA) [10] and linear subspace ranking hashing (LSRH) [14]. Ideally, hashing ensures that objects with high similarity renders higher probability of collision in the hashed domain; in contrast, the objects are far apart from each other result a lower probability of hash collision. Therefore, hashing is able to retain or even boost the matching performance in subspace.

In literature, Klare et al. [18] proposed a VIS-NIR face recognition method based on linear discriminant analysis (LDA) [19]. A common feature-based representation for both NIR images and VIS images are selected, then projected onto random subspace. The matching process is implemented with projected features using sparse representation classification. Klare et al. [20] further developed a kernel prototype similarities method, which is similar with earlier method [18] but requires no common features. Kernel prototype similarities method is tested on visible-NIR, visible-thermal and visible-sketch cases. Jin et al. [17] reported a coupled discriminative feature learning (CDFL) method. CDFL directly learns discriminative features from raw pixels. Then, the features are encoded and projected onto one common space with feature dimension reduction and matching accuracy improvement. Sharma et al. in [10] suggest that ideal inter-domain retrieval/classification requires that the samples from the same identity are as close as possible while those from different identities are far separated in the projected subspace. Moreover, Sharma et al. developed a cross-view face recognition system, namely generalized multiview analysis (GMA), and summarised the properties of a decent cross-view classification method, i.e. (i) Supervised, (ii) Generalizable, (iii) Multi-view, (iv) Efficient, (v) Kernelizable, (vi) Domain-Independent.

The proposed SPH is a ranking-based hashing method that rank correlation measures are taken into account whereas the aforementioned methods do not utilize the ordinary information [21]. Compared to the pairwise-based hashing methods [10], [17], [18], [20], ranking-based methods enjoy a superiority of performance as optimizing the ranking list can directly improve the quality and efficiency of nearest neighbor search [22]. From the viewpoint of loss function, SPH imposes not only the inter-domain but also intra-domain similarity measurement, which is rarely addressed in the literature. This theory that results in good performances can be found and justified in [10].

III. Preliminaries

In this section, we give a brief introduction for LSRH, in which the proposed SPH based upon. LSRH was originally meant for inter-domain (e.g. image-to-text and text-to-image) data retrieval.

A. Ranking-based hash function

Ranking-based hash function defines a family of hash function based on the max-order-statistics of feature projections

onto a K-dimensional linear subspace. The name of ranking-based hash function is originated from the ordinal of feature dimensions rather than the precise numerical values [14]. The ranking-based hash function $\mathbf{h}_*(\cdot)$ is defined as

$$\mathbf{h}(\mathbf{z};\mathbf{W}) = \arg\max_{1\le k\le K} \mathbf{w}_k^T\mathbf{z}, \tag{1}$$

where $\mathbf{z}$ is a set of D-dimensional feature vectors of samples and $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_K]^T \in \mathbb{R}^{K\times D}$ is a K-dimensional hashing matrix.

B. Formulation of LSRH

Loss function of LSRH is designed for only inter-domain matching purpose [14]. It would be part of our proposed method.

Assume that X and Y are face images from two domains. Let $\{\mathbf{x}_i\}_{i=1}^{N_X}$ and $\{\mathbf{y}_j\}_{j=1}^{N_Y}$ represent a set of feature vectors from each domain, where N_* denotes the number of samples in the respective domain. '$*$' is a place holder for X and Y. Let h_Z^i represent $\mathbf{h}_Z(\mathbf{z}_i;\mathbf{W}_Z)$, then (h_X^i, h_Y^j) indicates the K-ary ranking hashed codes of a inter-domain pair $(\mathbf{x}_i, \mathbf{y}_j)$. Then let π_{ij} be the probability of (h_X^i, h_Y^j) pair taking same value and s_{ij} be the similarity label (1 or 0) of $(\mathbf{x}_i, \mathbf{y}_j)$. Therefore, π_{ij} is expected close to 1 when s_{ij} is 1. Otherwise, π_{ij} is expected close to 0. The error function is defined as

$$\tilde{\ell}_{ij} = \begin{cases} 1-\pi_{ij}, & s_{ij}=1 \\ \lambda\pi_{ij}, & s_{ij}=0 \end{cases}, \tag{2}$$

where λ controls the relative penalty of false-positive pairs. Ranking-based hash function defined in (1) could be formulated and approximated using the softmax function as

$$\mathbf{h}(\mathbf{z};\mathbf{W}) \approx \sigma(\mathbf{W}\mathbf{z}), \tag{3}$$

where the function $\sigma(\mathbf{x})_j$ represents the j^{th} dimension of the output vector and is defined as:

$$\sigma(\mathbf{x})_j = \frac{e^{\alpha x^j}}{\sum_{k=1}^{K} e^{\alpha x^k}}, \tag{4}$$

where $j \in \{1, 2 \dots K\}$ and α controls the smoothness. Let $\mathbf{p}_i$ represent $\sigma(\mathbf{W}_X\mathbf{x}_i)$ and $\mathbf{q}_j$ represent $\sigma(\mathbf{W}_Y\mathbf{y}_j)$ as the softmax vectors. Therefore,

$$\pi_{ij} = \mathbf{p}_i^T\mathbf{q}_j. \tag{5}$$

From the above formulations, the overall loss function of LSRH is:

$$\begin{aligned}\tilde{L}_\alpha(\mathbf{W}_X,\mathbf{W}_Y) &= \sum_{s_{ij}=1}(1-\pi_{ij}) + \sum_{s_{ij}=0}\lambda\pi_{ij} \\ &= \sum_{s_{ij}\in\mathbf{S}} a_{ij}\mathbf{p}_i^T\mathbf{q}_j + const, \\ &= trace(\mathbf{P}\mathbf{A}\mathbf{Q}^T) + const\end{aligned} \tag{6}$$

where $\mathbf{P} = \left[\mathbf{p}_1 \dots \mathbf{p}_{N_X}\right]$ and $\mathbf{Q} = \left[\mathbf{q}_1 \dots \mathbf{q}_{N_Y}\right]$ are K-by-N_* matrix with softmax vectors in each column, and a matrix $\mathbf{A}$ with entries of N_X-by-N_Y is defined as

$$a_{ij} = \lambda - (\lambda+1)s_{ij}. \tag{7}$$

const is a fixed value, which equals to the amount of positive similarity labels $s_{ij}=1$. The training process aims to find

$$\min_{\mathbf{W}_X,\mathbf{W}_Y} \tilde{L}_\alpha = trace(\mathbf{P}\mathbf{A}\mathbf{Q}^T). \tag{8}$$

IV. THE SUBSPACE PROJECTION HASHING

LSRH is proposed to measure the similarity of data from different modalities, e.g. inter-domain for image and text retrieval. It merely takes care of the inter-domain similarity on the subspace (i.e. visible-thermal) [14]. However, Sharma et al. [10] revealed that the performance can be improved if intra-domain similarity (i.e. visible-visible and thermal-thermal) can be taken into account. Therefore, a new hashing framework SPH is proposed to capture our goal, i.e. inter-domain and intra-domain similarity measurement, as Fig. 1 illustrates. SPH is a ranking-based hashing and loss function is optimized for both inter-domain and intra-domain matching.

First, the singular ranking-based hash function $\mathbf{h}_*(\cdot)$ of SPH is defined as

$$\mathbf{h}_X(\mathbf{x};\mathbf{W}_X) = \arg\max_{1\le k\le K} \mathbf{w}_{Xk}^T\mathbf{x}, \tag{9}$$

$$\mathbf{h}_Y(\mathbf{y};\mathbf{W}_Y) = \arg\max_{1\le k\le K} \mathbf{w}_{Yk}^T\mathbf{y}, \tag{10}$$

where $\mathbf{W}_* = [\mathbf{w}_{*1} \dots \mathbf{w}_{*K}]^T \in \mathbb{R}^{K\times D_*}$ is a K-dimensional hashing matrix. The hashed code length is defined as L, which means L duplicates of $\mathbf{h}_*(\cdot)$. With a well-designed hashing/projection matrix $\mathbf{W}_*$, two sets of L-dimensional hashed codes for two domains can be generated by:

$$\mathbf{H}_* = \left\{\mathbf{h}_*^l\right\}_{l=1}^{L}. \tag{11}$$

Next, we re-formulate (6) as a part of SPH loss, to preserve the inter-domain similarity.

$$\begin{aligned}\tilde{L}_{cross} &= \sum_{s_{x_iy_j}\in\mathbf{S}_{XY}} a_{x_iy_j}\mathbf{p}_i^T\mathbf{q}_j + const_{XY} \\ &= trace(\mathbf{P}\mathbf{A}_{XY}\mathbf{Q}^T) + const_{XY}\end{aligned}, \tag{12}$$

where $\mathbf{A}_{XY}$ is defined as

$$a_{x_iy_j} = \lambda - (\lambda+1)s_{x_iy_j}. \tag{13}$$

$s_{x_iy_j}$ is the similarity label of $(\mathbf{x}_i, \mathbf{y}_j)$ and $const_{XY}$ is a fixed value of the amount of positive similarity labels $s_{x_iy_j}=1$.

Next, the intra-domain similarity preserving loss function is introduced from (12) and (13). The intra-domain similarity preserving loss is defined as:

$$\begin{aligned}\tilde{L}_{single} = \frac{1}{2}(&trace(\mathbf{P}\mathbf{A}_{XX}\mathbf{P}^T) + const_{XX} \\ &+trace(\mathbf{Q}\mathbf{A}_{YY}\mathbf{Q}^T) + const_{YY}),\end{aligned} \tag{14}$$

where $trace(\mathbf{P}\mathbf{A}_{XX}\mathbf{P}^T)$ and $trace(\mathbf{Q}\mathbf{A}_{YY}\mathbf{Q}^T)$ represent intra-X and intra-Y domain, respectively, where $\mathbf{A}_{XX}$ is defined as

$$a_{x_ix_j} = \lambda - (\lambda+1)s_{x_ix_j}, \tag{15}$$

and $\mathbf{A}_{YY}$ is defined as

$$a_{y_iy_j} = \lambda - (\lambda+1)s_{y_iy_j}. \tag{16}$$

$s_{x_i x_j}$ and $s_{y_i y_j}$ are similarity label of X and Y domain, respectively. $const_{XX}$ and $const_{YY}$ are fixed values of the amount of positive similarity labels $s_{x_i x_j} = 1$ and $s_{y_i y_j} = 1$, respectively.

To optimize projection matrix, an additional regularization term, i.e. pairwise distance is introduced to the original loss function. The pairwise distance penalty is computed based on pairwise distance between each hashed code on a common subspace. We define pairwise distance of a set of vectors $\mathbf{Z} = [\mathbf{z}_1 \dots \mathbf{z}_N]$ as

$$\mathbf{D}(\mathbf{Z}) = \{dist(\mathbf{z}_i, \mathbf{z}_j)\}, \forall\ i < j\ and\ i, j \in N, \tag{17}$$

where $\mathbf{D}(\mathbf{Z}) \in \mathbb{R}^{1 \times C_N^2}$. C_N^2 is the mathematical combination calculation. It represents the number of different combinations for selecting two items from N items collection, ignoring the order of selection. Let $\mathbf{Z} = [\mathbf{h}_X(\mathbf{x}; \mathbf{W}_X), \mathbf{h}_Y(\mathbf{y}; \mathbf{W}_Y)]$. Therefore, all pairwise distances of hashed codes are $\mathbf{D}(\mathbf{Z}) \in \mathbb{R}^{1 \times C_{N_X+N_Y}^2}$. $\mathbf{D}$ is normalized into (0,1). Since we expect to maintain larger pairwise distance $\mathbf{D}(\mathbf{Z})$, the loss function is defined as

$$\tilde{L}_{pdist} = 1 - \frac{\sum_{i=1}^{C_{N_X+N_Y}^2} \mathbf{D}_i(\mathbf{Z})}{C_{N_X+N_Y}^2}. \tag{18}$$

The overall loss function of SPH is to fuse (12), (14) and (18), as

$$\tilde{L}_{SPH}(\mathbf{W}_X, \mathbf{W}_Y) = \tilde{L}_{cross} + \beta_1 \tilde{L}_{single} + \beta_2 \tilde{L}_{pdist}. \tag{19}$$

β_1 and β_2 are the coefficients of new loss functions. We can observe that when $\beta_1 = 0$, the overall loss function is optimal for inter-domain matching performance and when $\beta_1 = \beta_2 = 0$, SPH is identical to LSRH. Thus, SPH is a generalized form of LSRH. The outputs of SPH are two sets of L-dimensional hashing functions/project matrices, i.e. $(\mathbf{W}_X, \mathbf{W}_Y)$. The general algorithm is indicated in Algorithm 1.

Algorithm 1: Subspace Projection Hashing

Input : Face features $\mathbf{X}$, $\mathbf{Y}$ and similarity labels $\mathbf{S}$, subspace dimension K, hashed code length L.
Output: K-dimensional hashing matrix $\mathbf{W}_X$ and $\mathbf{W}_Y$

Initialization: Set $\mathbf{W}_X$ and $\mathbf{W}_Y$ to random values from Gaussian distribution
repeat
  Randomly select a training batch X_b and Y_b, obtain the batchwise label matrix $S_{X_b X_b}$, $S_{Y_b Y_b}$ and $S_{X_b Y_b}$ accordingly;
  Update $\mathbf{W}_X$ and $\mathbf{W}_Y$ according to mini-batch gradient decent method in [14].
until *Convergence*

V. EXPERIMENT

We conducted a series of experiments to demonstrate the feasibility of the proposed SPH for cross-spectrum FR. We have made our testing data and source code publicly available for reproducing our results (https://github.com/azrealwang/sph).

A. Dataset and matching protocol

Wiki dataset [23] is used for inter-domain performance evaluation in LSRH [14]. To ensure a fair comparison, Wiki is also adopted to validate the effectiveness of the proposed loss function in SPH. Wiki dataset is based on Wikipedia's "featured articles" and involves 2,866 image-text pairs with semantic labels of 10 categories. Each image is represented as a 128-dimensions feature vector and each text document is represented as a 1000-dimensions feature vector. We use mean average precision (mAP) as a performance indicator, to evaluate recognition accuracy.

EURECOM VIS-TH paired face dataset is a newly released dataset dedicated for inter-domain FR, e.g. visible-thermal. The samples of facial images are illustrated in Fig. 2. EURECOM VIS-TH paired face dataset involves 2,100 face images, including 50 identities with 21 visible-thermal pairs for each identity [24]. Face features are extracted as a 512-dimensions feature vector with public software InsightFace [25]. Rank-1 recognition rate and equal error rate (EER) are adopted to measure the recognition performance.

TDFace dataset includes more than 100 subjects of visible-NIR face pairs [26]. Fig. 3 illustrates the samples in this dataset. To increase the diversity of experiments, the face features are presented as a 512-dimensions feature vector with FaceNet [1]. Due to the failure of face image alignment, some of the face images are excluded for experiment and eventually 3,180 images, including 106 identities with 15 visible-NIR pairs. It is evaluated with Rank-1 rate.

B. Training strategy

To generate the training-set and testing-set, Weinberger et al.'s protocol [27] is adopted in this paper. For EURECOM VIS-TH paired face, 15 images from each identity are randomly selected as training-set and the remaining 6 images are used as testing-set. For TDFace, 11 images of each identity are selected randomly as training-set, while the remaining 4 images are used for testing-set. To ensure the generalizability

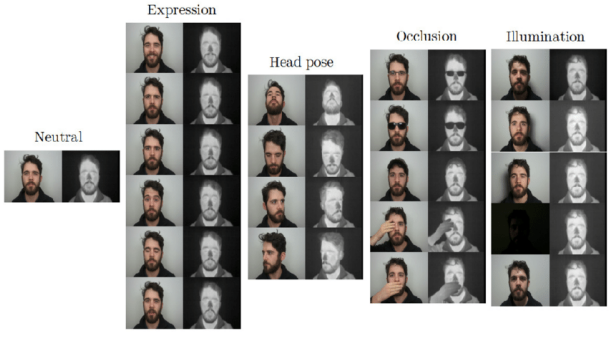


Fig. 2. Samples in EURECOM VIS-TH paired face dataset [24].

Fig. 3. Samples in TDFACE dataset [26].

Algorithm 2: Subspace Projection Hashing Incremental Learning

Input : Historical selected data $\mathbf{X}_h$, $\mathbf{Y}_h$ and $\mathbf{S}_h$; historical hashing matrix $\mathbf{W}_{Xh}$ and $\mathbf{W}_{Yh}$; unseen data $\mathbf{X}_u$, $\mathbf{Y}_u$ and $\mathbf{S}_u$.

Output: Hashing matrix $\mathbf{W}_X$ and $\mathbf{W}_Y$; updated historical selected data $\mathbf{X}_h$, $\mathbf{Y}_h$ and $\mathbf{S}_h$

Set $\mathbf{W}_X$ and $\mathbf{W}_Y$ with historical hashing matrix $\mathbf{W}_{Xh}$ and $\mathbf{W}_{Yh}$

Set $\mathbf{X}$ with combining $\mathbf{X}_h$ and $\mathbf{X}_u$; set $\mathbf{Y}$ with combining $\mathbf{Y}_h$ and $\mathbf{Y}_u$; set $\mathbf{S}$ with combining $\mathbf{S}_h$ and $\mathbf{S}_u$

Update $\mathbf{X}_h$, $\mathbf{Y}_h$ and $\mathbf{S}_h$ with randomly selecting one image of each identity from $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{S}$

repeat

  Using mini-batch gradient decent method in Algorithm 1

until *training is terminated with stop condition* γ

of experiment results, all images selection for each time experiment is random and the experiments are repeated for 20 times (Table I). The average recognition accuracy is recorded in Table II - V.

In reality, unseen identities (identities not used in training) are commonly expected in a system, which will lead to a poor recognition performance [10], [28]. Though a new model can be re-trained by including the unseen identities and the historical training data, the process of re-train can be time costly and computation expensive. Hence an adaptive incremental learning strategy (SPH-IL) is designed to update the model incrementally with the unseen identities and part of the history training data, as described in Algorithm 2. Specifically, the new training-set is generated by combining the unseen dataset and one randomly selected image of each identity in the previous training-set. The weight in the pre-trained model is used as the initial weight in the re-train process. The re-train process is terminated with a stop condition γ, when the unseen identities recognition accuracy and the pre-trained model matching accuracy is close.

Conventionally, two independent rounds of training consume much time. The time consumption can be largely reduced by applying incremental learning where only small number of images (unseen identities and partial historical data) is required for re-train training-set and re-train process is terminated shortly. Therefore, the time of re-train would be no longer a concern. Experiment results in section V-E demonstrates the efficiency and effectiveness of the incremental learning as opposed to the conventional training.

TABLE I
PARAMETERS USED IN EXPERIMENTS FOR EACH DATASET.

Parameter	WIKI	EURECOM	TDFACE
Duplicate of experiment		20	
Hashed code lengths (L)	128	128	128
Subspace dimensions (K)	8	16	16
Smoothness of (4) (α)	8	2	2
Penalty of (2) (λ)	1	0.2	0.1
Influence index of $\tilde{L}_{single}$ (β_1)	0	0.5	0.5
Influence index of $\tilde{L}_{pdist}$ (β_2)	0.5	0.5	0.5
Stop condition (γ)	N/A	0.22	N/A

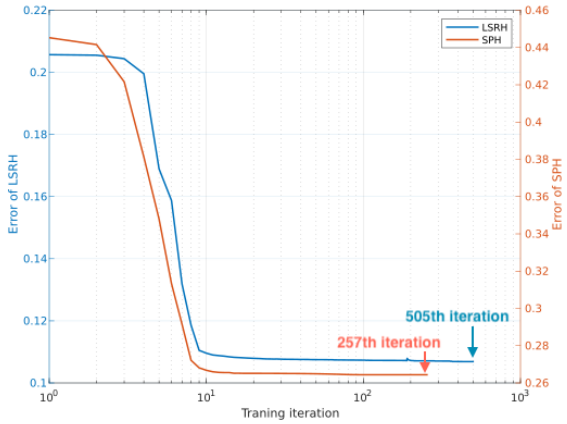


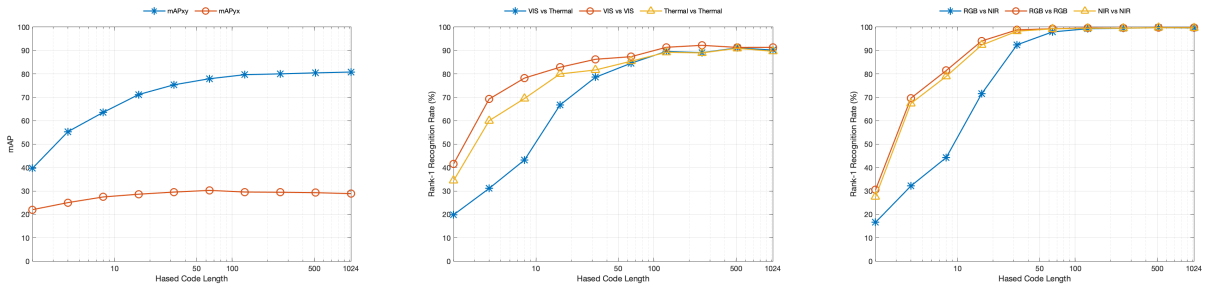
Fig. 4. Gradient decent of LSRH and proposed SPH for wiki dataset. Training terminated when no significant error decreased between consecutive iterations, in our case it's less than $2^{-52} \approx 2.22e^{-16}$ (i.e. eps in Matlab). Note that error of SPH is higher than LSRH is expected due to the additional loss $\tilde{L}_{single}$ and $\tilde{L}_{pdist}$ in SPH.

C. Parameters

The experimental parameters used in this paper are given in Table I. Note that all parameter values in Table I are optimal for best matching performance. Specifically, stop condition is only adopted in visible-thermal experiments (EURECOM VIS-TH paired face) for incremental learning testing. We conducted an experiment to observe the performance with respect to the length of the hashed code. Fig. 5 illustrates that the Rank-1 rate increases when the length of hashed code, L increases and levels off when L reaches 128 dimensions. The increment of rank-1 rate is insignificant when L is more than 128, whereas the computational complexity becomes a concern.

D. Comparison between LSRH and SPH on image-text retrieval

With the new loss function employed, we investigated the performance of SPH over LSRH for inter-domain (e.g. image-text) retrieval. The experiment is conducted using Wiki dataset



(a) Wiki (b) EURECOM VIS-TH paired face (c) TDFACE

Fig. 5. Recognition accuracy vs hashed code length for different datasets.

TABLE II
COMPARISON BETWEEN SPH AND LSRH.

mAP (%)	Text query image	Image query text
Proposed SPH	**79.66**	**29.55**
LSRH [14]	73.53	28.72

[14], [23] to ensure a fair comparison. The results shown in Table II indicate that the proposed SPH outperforms its predecessor LSRH. We examined the efficiency of convergence with the new loss function. The loss gradient descent of each iteration is plotted in Fig. 4. It is observed that the loss gradient of SPH only costs around half iterations to convergence, respect to LSRH. Thus, SPH is more training iteration efficient than LSRH, which means, SPH is more efficient to obtain the optimized hashing/projection matrices with an equal number of iterations. In addition, the figure indicates that for both SPH and LSRH, loss error has achieved a satisfied convergence after 100 training iterations, so our experiment setting of maximum iterations is 100 to reduce unnecessary computational cost. Note that training time of SPH is longer than LSRH since the proposed loss function increases the computational cost for each iteration. The time spent in seconds for the two lines in Fig. 4 is SPH: 62 seconds for 257 iterations vs LSRH: 45 seconds for 505 iterations, with MATLABVer. 2019b, 2.7 GHz Dual-Core Intel Core i5 CPU and 1867MHz 8GB RAM.

E. Performance of SPH on cross-spectrum FR

To evaluate the performance of SPH on VIS-TH FR, experiments are carried out on EURECOM VIS-TH paired face dataset with matching inter-domain (VIS-TH) and intra-domain (VIS-VIS and TH-TH) faces. Table III presents the Rank-1 recognition rate and matching EER between visible-thermal, visible-visible and thermal-thermal faces. We observed from Table III that the deep model underperformed for visible-thermal FR, e.g. InsightFace achieved Rank-1 recognition rate and EER for visible-thermal are 23.95% and 28.20% respectively. This is expected that the majority of the texture and edge information are lost due to the large distinction between visible and thermal spectrum as outlined in [7].

TABLE III
ACCURACY OF VISIBLE-THERMAL FR FOR EURECOM DATASET.

Rank-1 (%)	VIS vs TH	VIS vs VIS	TH vs TH
InsightFace	23.95	91.43	82.52
NMDSH [4]	26.58	**94.60**	83.83
CRN [29]	54.23	91.65	18.00
LSRH [14]	88.92	91.08	88.88
Proposed SPH	**89.62**	91.38	**89.25**
EER (%)	VIS vs TH	VIS vs VIS	TH vs TH
InsightFace	28.20	9.17	11.86
NMDSH [4]	27.88	**5.48**	9.75
LSRH [14]	7.56	7.28	7.78
Proposed SPH	**7.06**	6.95	**7.21**

TABLE IV
RANK-1 RECOGNITION RATE OF VISIBLE-NIR FR FOR TDFACE DATASET.

Rank-1 (%)	VIS vs NIR	VIS vs VIS	NIR vs NIR
FaceNet	95.13	**99.88**	**99.68**
Circular with HOG [30]	96.82	Not applicable	Not applicable
LSRH [14]	99.06	99.68	99.59
Proposed SPH	**99.30**	99.67	99.49

In contrast, SPH achieved the accuracy of **89.62%** Rank-1 rate and **7.06%** EER respectively, which suggests that the proposed loss is effective for inter-domain matching (VIS-TH). Additionally, two cross-spectrum (VIS-TH) FR methods, i.e. non-linear multi-dimensional spectral hashing (NMDSH) [4] and cascaded refinement network (CRN) [29] are chosen for performance comparison, because both NMDSH and CRN employed EURECOM VIS-TH paired face dataset that is the same dataset the proposed SPH used. We note that the experimental protocol used in CRN is that 45 subjects are for training and 5 subjects are for testing respectively. Although, this protocol is not exactly same to the protocol SPH used, it is the most similar protocol to ours (section V-B). It can be observed that the proposed SPH attains supreme accuracy for VIS-TH and TH-TH FR over NMDSH and CRN. Furthermore, while the recognition accuracy of SPH outperforms

TABLE V
PERFORMANCE OF INCREMENTAL LEARNING.

Scenarios	Experiment settings			Experiment results		
	Training-set	Gallery	Probe	VIS vs TH	VIS vs VIS	TH vs TH
SPH without unseen identities	Randomly selected 15 vis-th image pairs of each identity × 50 identities	Remaining 6 images of each identity in vis or th domain × 50 identities	Remaining 6 images of each identity in vis or th domain × 50 identities	89.62	91.38	89.25
SPH with unseen identities	Randomly selected 15 vis-th image pairs of each identity × 40 identities	Randomly selected 6 images of each identity in vis or th domain × 10 unseen identities	Randomly selected 6 images of each identity in vis or th domain × 10 unseen identities	1.33	35.92	22.58
		Remaining 6 images of each identity in vis or th domain × 50 (include 10 unseen) identities	Remaining 6 images of each identity in vis or th domain × 50 (include 10 unseen) identities	71.13	79.83	75.33
SPH-IL with unseen identities	*Initial:* Randomly selected 15 vis-th image pairs of each identity × 40 identities *Incremental:* Combine one randomly selected image of each identity in initial training-set, and randomly selected 15 vis-th image pairs of each identity × 10 unseen identities	Remaining 6 images of each identity in vis or th domain × 50 (include 10 unseen) identities	Remaining 6 images of each identity in vis or th domain × 50 (include 10 unseen) identities	86.67	90.65	87.38

the recognition accuracy of other schemes under inter-domain (VIS-TH), simultaneously, SPH also achieved the comparable performance under intra-domain, i.e. 91.38% for VIS-VIS and 89.25% for TH-TH. These results verified the theoretical justification of the proposed loss described in Section IV, where the proposed loss optimizes both inter-domain and intra-domain similarity preservation, and converges with a more optimized hashing matrix with same number of iterations.

For VIS-NIR, we conducted the experiments to examine the performance on inter-domain (VIS-NIR) and intra-domain (VIS-VIS and NIR-NIR) using TDFace dataset. Table IV tabulates the Rank-1 recognition rate between visible-NIR, visible-NIR and NIR-NIR faces. We can observe that despite FaceNet achieved a 95.13% of Rank-1 recognition rate for inter-domain (VIS-NIR), SPH even improved the Rank-1 recognition rate to **99.30%**. This result is comparable to the state-of-the-art deep model for inter-domain (VIS-NIR) [6]. In addition, we compared the performance of SPH with a recently proposed method, namely circular with histogram of oriented gradients (HOG) [30], which also used TDFace dataset for VIS-NIR FR. The experimental results shown in Table IV indicate that the proposed SPH achieved a higher accuracy over HOG. Furthermore, similarly with intra-domain experiment results of EURECOM VIS-TH paired face dataset, intra-domain matching performance in TDface dataset is also equivalently comparable with other methods. These results confirms that SPH is effective on VIS-NIR FR.

As mentioned in section V-B, the poor recognition performance for unseen identities is commonly expected in reality. Unseen identities refer to the identities which do not participate in the training process. To tackle with this issue, we designed an adaptive incremental training strategy (SPH-IL) and examined the effectiveness under this scenario. Table V provides the performance comparisons between incremental learning strategy and conventional learning strategy for unseen identities scenario. We can observe that the recognition performance under the unseen identities scenario is completely jeopardized (1.33%, 35.92% and 22.58% of Rank-1 rate for VIS-TH, VIS-VIS, and TH-TH). In contrast, by employing the incremental learning strategy, the recognition performances can gain 86.67%, 90.65%, and 87.38% Rank-1 rate for VIS-TH, VIS-VIS, and TH-TH respectively. Overall, the incremental learning strategy scarified an insignificant performance reduction (i.e. -2.95% Rank-1 rate from Table V) but gaining the capability of dealing with unseen identities without requiring overwhelming computational overload.

F. Discussion

From the experimental results above, we can summarise that the recognition accuracy of SPH performs well under inter-domain scenario, i.e. visible-thermal and visible-NIR. This is attributed to the proposed loss that generates a more optimized projection matrix with same iterations. On the other hand, the recognition accuracy of SPH is comparable (no significant performance reduction) to the state-of-the-art deep models under the intra-domain scenario (i.e. VIS-VIS, TH-TH and NIR-NIR). Note that the deep models are specialized for intra-domain FR. SPH could be comparable to deep models is due to the proposed loss that additionally preserves the intra-domain similarity.

VI. CONCLUSION

In this paper, we proposed a hashing method, namely SPH that is extended from LSRH. Raw face features are hashed to

less dimensional integer codes, which increases matching and searching speed. We specially designed a new loss function that preserves the inter-domain and intra-domain similarity. Simultaneously, the pairwise distance as penalty is included in the loss function to optimize projection matrix. Lastly, we introduced the incremental learning procedure to tackle the unseen identities problem. The experiment results on Wiki dataset suggest that the new loss function achieved better performance over LSRH and demonstrated less iterations to convergence as well. For cross-spectrum FR, the experiments on VIS-TH and VIS-NIR indicate that SPH outperforms LSRH and deep learning approaches (e.g. InsightFace and FaceNet). Moreover, the unseen identities problem has been largely alleviated by the incremental learning procedure as indicted by the experimental results.

In the future work, we are keen to investigate the feasibility of SPH applying on other modalities such as visible-sketch FR and 2D-3D model retrieval. We will also explore other possible directions to design more sophisticated loss functions such as modifying or fusing loss for cross-spectrum FR.

ACKNOWLEDGMENT

This work was supported by the EUropean Union COST Action CA16101, University of Sassari, fondo di Ateneo per la ricerca 2019, Italian Ministry for Research, special research project SPADA and Fundamental Research Grant Scheme (FRGS/1/2018/ICT02/MUSM/03/3) and NVIDIA Corporation donation of the Titan Xp GPU.

REFERENCES

[1] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 815-823.

[2] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, 2010, pp. 2567-2573.

[3] A. K. Jain, B. Klare and U. Park, "Face matching and retrieval in forensics applications," *IEEE Multimedia*, vol. 19, no. 1, p. 20-28, 2012.

[4] X. Dong, K. Wong, Z. Jin and J.-L. Dugelay, "A secure visual-thermal fused face recognition system based on non-linear hashing," in *IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, Kuala Lumpur, Malaysia, Malaysia, 2019, pp. 1-6.

[5] K. Mallat, N. Damer, F. Boutros, A. Kuijper and J.-L. Dugelay, "Cross-spectrum thermal to visible face recognition based on cascaded image synthesis," in *International Conference on Biometrics (ICB)*, Crete, Greece, Greece, 2019, pp. 1-8.

[6] Z. Deng, X. Peng, Z. Li and Y. Qiao, "Mutual component convolutional neural networks for heterogeneous face recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 3102-3114, June 2019.

[7] T. Zhang, A. Wiliem, S. Yang and B. C. Lovell, "TV-GAN: Generative adversarial network based thermal to visible face recognition," in *International Conference on Biometrics (ICB)*, Gold Coast, QLD, Australia, 2018, pp. 174-181.

[8] H. Zhang, B. S. Riggan, S. Hu, N. J. Short and V. M. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 845-862, 2019.

[9] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, 2006.

[10] A. Sharma, A. Kumar, H. Daume and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, 2012, pp. 2160-2167.

[11] D. R. Hardoon, S. Szedmak and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639-2664, 2004.

[12] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247-1283, 2000.

[13] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, 2011, pp. 593-600.

[14] K. Li, G.-j. Qi, J. Ye and K. A. Hua, "Linear subspace ranking hashing for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1825 - 1838, 2016.

[15] W. Liu, C. Mu, S. Kumar and S.F. Chang, "Discrete graph hashing," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 3419-3427,.

[16] H. Zhu, M. Long, J. Wang and Y. Cao, "Deep hashing network for efficient similarity retrieval," in *Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, Phoenix, Arizona, USA, 2016.

[17] Y. Jin, J. Lu and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 640-652, March 2015.

[18] B. Klare and A. K. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *2010 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, 2010, pp. 1513-1516.

[19] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *International Conference on Biometrics (ICB)*, Springer, Berlin, Heidelberg, 2009, pp. 209-218.

[20] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1410-1422, June 2013.

[21] M. Melucci, "On rank correlation in information retrieval evaluation," in *ACM SIGIR Forum*, vol. 41, no. 1, pp. 18-33, 2007.

[22] J. Wang, W. Liu, A. X. Sun and Y. Jiang, "Learning hash codes with listwise supervision," in *2013 IEEE International Conference on Computer Vision (ICCV)*, Sydney, NSW, 2013, pp. 3032-3039.

[23] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM International Conference on Multimedia (ACMMM)*, Firenze, Italy, 2010, pp. 251-260.

[24] K. Mallat and J.-L. Dugelay, "A benchmark database of visible and thermal paired face images across multiple variations," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2018, pp. 1-5.

[25] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690-4699.

[26] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. Rao, A. Kaszowska, H. Taylor, A. Samani and X. Yuan, "A comprehensive database for benchmarking imaging systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[27] K. Q. Weinberger, J. Blitzer and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems (NIPS)*, 2006, pp. 1473-1480.

[28] F. Li and H. Wechsler, "Open set face recognition using transduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1686-1697, 2005.

[29] K. Mallat, N. Damer, F. Boutros, and J. L. Dugelay, "Robust Face Authentication Based on Dynamic Quality-weighted Comparison of Visible and Thermal-to-visible images to Visible Enrollments," *In 2019 22th International Conference on Information Fusion (FUSION)*, pp. 1-8, July. 2019.

[30] S. Rajeev, S. K. KM, Q. Wan, K. Panetta, and S. S. Agaian, "Illumination invariant NIR face recognition using directional visibility," *Electronic Imaging*, vol. 2019, no. 11, pp. 273-1, 2019.