The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification

Mang Ye, Xiangyuan Lan, Jiawei Li, Pong C. Yuen*
Department of Computer Science, Hong Kong Baptist University
{mangye,jwli,pcyuen}@comp.hkbu.edu.hk, xiangyuanlan@life.hkbu.edu.hk

Abstract

Person re-identification is widely studied in visible spectrum, where all the person images are captured by visible cameras. However, visible cameras may not capture valid appearance information under poor illumination conditions, e.g, at night. In this case, thermal camera is superior since it is less dependent on the lighting by using infrared light to capture the human body. To this end, this paper investigates a cross-modal re-identification problem, namely visible-thermal person re-identification (VT-REID). Existing cross-modal matching methods mainly focus on modeling the cross-modality discrepancy, while VT-REID also suffers from cross-view variations caused by different camera views. Therefore, we propose a hierarchical cross-modality matching model by jointly optimizing the modality-specific and modality-shared metrics. The modality-specific metrics transform two heterogenous modalities into a consistent space that modality-shared metric can be subsequently learnt. Meanwhile, the modality-specific metric compacts features of the same person within each modality to handle the large intra-modality intra-person variations (e.g. viewpoints, pose). Additionally, an improved two-stream CNN network is presented to learn the multi-modality sharable feature representations. Identity loss and contrastive loss are integrated to enhance the discriminability and modality-invariance with partially shared layer parameters. Extensive experiments illustrate the effectiveness and robustness of the proposed method.

Introduction

Person re-identification (REID) addresses the problem of matching different persons across disjoint camera views (Zheng, Yang, and Tian 2017; Wang et al. 2016b). It has gained much attention in recent computer vision communities due to its importance in intelligent video surveillance (Lan, Zhang, and Yuen 2016; Lan, Yuen, and Chellappa 2017; Lan et al. 2015). Most of current re-id methods are focusing on visible images, i.e., given a probe image/video and match it against a set of gallery images/videos. Under this visible domain to model the large cross-camera variations, encouraging results have been achieved (Ye et al. 2017) by 1) learning or designing representative features for visual appearance of persons, and 2) learning proper similarity measurements on the visual feature vectors.

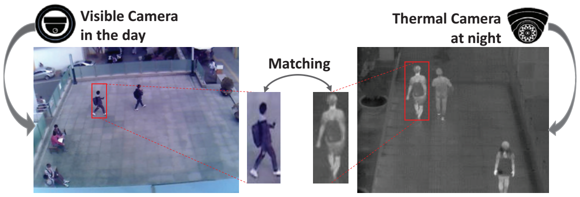

Figure 1: Illustration about the visible thermal person re-identification (VT-REID). Person images from different modalities should be matched.

However, the visible camera may not be able to capture all the appearance information for person re-identification, especially under poor illumination conditions. In these cases, thermal camera is superior since it is less dependent on the lighting by using infrared light to capture the human body. Therefore, thermal cameras can be introduced for robust person re-identification under the low illumination environments. As shown in Fig.1, the query images are captured by visible cameras during the daytime, while gallery images are obtained from thermal cameras at night. In this paper, we name this cross-modal person re-identification problem as Visible-Thermal person Re-identification (VT-REID).

As far as we know, two pioneer works have investigated the thermal image based re-identification problem. In 2013, Mogelmose *et al.* (Mogelmose et al. 2013) propose a tri-modal re-identification model by integrating three modalities features (rgb, depth and thermal) to improve the re-identification performance. Similarly, Nguyen (Nguyen et al. 2017) introduces a person recognition system by combining the visible and thermal image features. Both of them try to fuse the thermal modality information to improve the traditional visible image based re-identification. Different from their works, we aim at investigating this cross-modal re-identification problem, thus need to bridge the gap between these two heterogeneous modalities.

Existing cross-modal re-identification methods (Li, Xiao, and et al. 2017) mainly focus on bridging the gap between the visual and text domains, which cannot be directly applied for VT-REID. Related cross-modal matching problem has been widely studied in many heterogenous recognition applications, such as thermal-visible face recognition, face sketch recognition and so on. Usually, they mainly focus

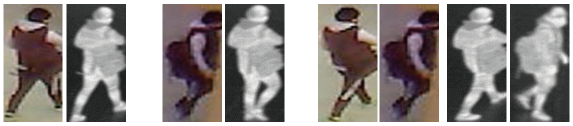

Figure 2: Different variations for VT-REID. All images are the same person. Left: *Cross-modality discrepancy*. Middle: *Cross-view variations*. Right: *Intra-modality intra-person variations*.

on modeling the *cross-modality discrepancy* (Sarfraz and Stiefelhagen 2017), arisen from their different reflective visible spectrums and sensed emissivities of visible and thermal cameras. However, VT-REID also suffers from large *cross-view variations* caused by different camera views as illustrated in Fig. 2. Besides that, even within the same modality, the appearance of the person images can change dramatically due to the pose and viewpoint variations (Ye et al. 2016; 2015b), we name it as *intra-modality intra-person variations*. Therefore, it is important to develop a discriminative model to simultaneously handle all the variations for VT-REID.

To address the aforementioned problems, this paper presents a novel hierarchical cross-modality matching model for VT-REID. We propose to jointly optimize the modality-specific and modality-shared metrics. Two different modality-specific metrics transform two heterogenous modalities into a consistent space that modality-shared metric can be subsequently learnt, thus could handle the *cross-view variation* problem. Specifically, modality-specific metric compacts the distances among the images of the same person within each modality, thus could also reduce the *intra-modality intra-person variations*. The modality-shared metric aims to learn a projection that could distinguish different persons across two heterogenous modalities, which tackles the *cross-modality discrepancy* problem. Apart from the matching model learning, an improved two-stream CNN network is introduced to learn deep feature representations for VT-REID. Discriminative multi-modality sharable feature representations for person images in two heterogenous modalities are achieved by integrating identity loss and contrastive loss. The identity loss aims to model domain-specific information to distinguish different persons within each modality. The contrastive loss bridges the gap between two heterogenous modalities and enhances the modality-invariance of the learnt representation.

The main contributions can be summarized as follows:

- A new research issue about the visible-thermal based person re-identification is investigated in this paper, which is important for practical surveillance applications.
- A novel hierarchical cross-modality matching model for VT-REID is proposed, which could simultaneously handle both cross-modality discrepancy and cross-view variations, as well as intra-modality intra-person variations.
- An improved two-stream CNN network is presented to learn the deep multi-modality sharable feature representations.

Related Work

In this section, some prior related works about person re-identification and other cross-modal recognition (retrieval) works are included for discussion.

Person Re-identification. A detailed survey about visible image based re-identification can be found in (Zheng, Yang, and Hauptmann 2016). Here we mainly focus on multi-modal fusion and cross-modal person re-identification.

Most existing multi-modal fusion based re-identification methods focus on RGB-D modules (Barbosa et al. 2012; Wu, Zheng, and Lai 2017), where the depth information captured by a depth camera could be integrated to the traditional RGB channels. In this manner, the re-identification performance is improved. Another most related work is presented in (Nguyen et al. 2017), they trained two general deep neural networks for thermal and visible modalities separately, and then fused the features for re-identification. All these works try to improve the re-id performance by fusing different modalities, while we focus on cross modal re-id problem.

Another two research works about cross-modal person re-identification are searching a specific person with text descriptions (Li, Xiao, and et al. 2017; Ye et al. 2015a), they aim to bridge the gap between the text descriptions and visual images by using a deep and shallow method, respectively. But their approaches cannot be directly adopted for visible thermal person re-identification problem.

Cross-modal Recognition (Retrieval). Cross-modal recognition problems are widely studied for heterogenous face recognition (He et al. 2017), text-to-image retrieval (Cao et al. 2017) and etc. We discuss these works from two perspectives: feature representations and matching models.

For feature representations, early researchers aim to design hand-crafted features (Sarfraz and Stiefelhagen 2017) and then adopting some machine learning techniques to bridge the gap between heterogenous modalities. However, hand-crafted features can not generalize well, and thereby it may not be able to preserve the discriminability in large scale applications. Comparatively, deep neural networks could learn robust feature representations with strong discriminability, and have shown promising results in many vision tasks (Cao et al. 2017). Therefore, it is extremely important to investigate to learn deep feature representations for VT-REID.

For learning matching models, dictionary learning[1] (Peng et al. 2017; Zhuang et al. 2013; Das, Mandal, and Biswas 2017) and metric learning (Huo et al. 2017; Wang et al. 2016a; 2017) are two widely used techniques. Recently, Sarfra *et al.* (Sarfraz and Stiefelhagen 2017) also propose a deep cross-modal matching algorithm with a two-layer nonlinear function to bridge the gap between the visual and thermal images. However, most existing cross-modal matching methods aim at addressing the cross-modality discrepancy, while our VT-REID problem also suffers from the large cross-view variations and intra-modality variations. The proposed method tries to address all these issues during the feature representation learning and hierarchical metric learning procedures.

[1] It can also be treated as a feature learning approach.

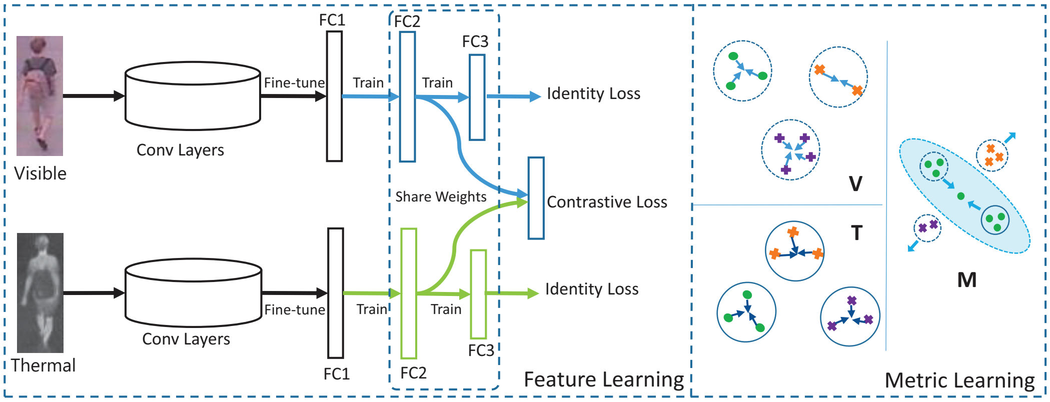

Figure 3: The proposed framework for VT-REID. It contains two stages, feature learning and metric learning. The former aims to learn multi-modality sharable feature representations by a two-stream CNN network constraining with identity loss and contrastive loss. The latter focuses on discriminative matching model learning with modality-specific (V, T) and modality-shared (M) metrics, respectively.

Proposed Method

This paper addresses the VT-REID via a two-stage framework as shown in Fig. 3, namely feature learning and metric learning. We propose a TwO-stream CNN NEtwork (TONE) to learn the multi-modality sharable feature representations for two heterogenous modalities, integrating contrastive loss to bridge the gap between two modalities and enhance the modality-invariance of the learnt representation. After that, a Hierarchical Cross-modality Metric Learning (HCML) method is introduced by jointly optimizing the modality-specific and modality-shared metrics. The modality-specific metrics (V and T) transform two heterogenous modalities into a consistent space, thus could handle the *cross-view variation* problem. Meanwhile, the modality-specific metrics can also address the *intra-modality intra-person variations*. The modality-shared metric (M) aims to minimize the *cross-modality discrepancy*, which could distinguish different persons from two heterogenous modalities.

Multi-Modality Sharable Feature Learning

We firstly introduces the TwO-stream CNN NEtwork (TONE) for deep multi-modality sharable feature learning, which contains a visible image stream and a thermal image stream. We utilize the pre-trained model on ImageNet for fine-tuning, since we only have limited data for training from the scratch. In our model, we adopt AlexNet[2] (Krizhevsky, Sutskever, and Hinton 2012) as our baseline network, which contains five convolutional layers ($conv1 \sim conv5$) and three fully connected layers ($fc1 \sim fc3$). We treat the output of $fc2$ (or $fc1$) in each stream as the feature representation for each person image, denoted by $\{x_i\}_{i=1}^{N_x}$ for visible images and and $\{z_j\}_{j=1}^{N_z}$ for thermal images, respectively. To learn the feature representations, two kinds of optimization objectives including two identity losses and one contrastive loss are introduced.

Identity Loss. The identity loss aims to learn the discriminative feature representations by using modality-specific information, which could distinguish differen persons within each modality. There are two streams CNN network in our architecture, we denote the learnt parameters for two different modalities as θ_1 and θ_2. The length of $fc3$ is defined by the number of persons (K), which is similar to many multi-class classification problems. After that, the cross-entropy loss is adopted for identity prediction. Specifically, the identity loss for visible images is denoted by

$$\hat{p_i} = softmax(\theta_1 \odot x_i) \tag{1}$$

$$\mathcal{L}_1(\theta_1, t, x) = -\sum_{i=1}^{K} y_i \log(\hat{p_i}) \tag{2}$$

where $\odot$ denotes the convolutional operation. t is the target identity (class), and $\hat{p_i}$ is the predicted probability, and y_i is the groundtruth probability vector where y_i for all i except $y_t = 1$. Similarly, we could get the identity loss for the thermal stream network, denoted by $\mathcal{L}_2(\theta_2, t, z)$.

Contrastive Loss. The contrastive loss tries to bridge the gap between two heterogenous modalities, which could also enhance the modality-invariance for the feature learning. Denote x, z as the outputs of $fc2$ layers of two streams, $l2$-normalization is firstly introduced before calculating the contrastive loss. Therefore, the contrastive loss is defined by

$$\mathcal{L}_3 = \frac{1}{2N} \sum_{n=1}^{N} y_n d_n^2 + (1 - y_n) \max\left(margin - d_n, 0\right)^2 \tag{3}$$

where N denotes the batch size, and $d_n = \|x_n - z_n\|_2$ represents the Euclidean distance of x_n and z_n. $margin$ is a pre-defined threshold, which is defined as 0.5 in all our experiments via cross-validation. y_n is the label of of two samples, indicating two images from two streams whether belong to the same person identity ($y_n = 1$) or not ($y_n = 0$). To balance the positive and negative affects, equal number of positive and negative pairs are selected for each batch.

[2] Other networks such as the VggNet, GoogLeNet and ResNet architectures can be configured without any limitation.

Feature Learning Optimization. By considering two identity losses and the contrastive loss, the overall loss function for our two stream CNN network is integrated as:

$$\min_{\theta_1,\theta_2,\theta_3} \mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \alpha\mathcal{L}_3 \tag{4}$$

where α is the trade-off between identity loss and contrastive loss. The optimization model can be solved by using existing deep learning toolboxes. By optimizing (4), we can successfully bridge the gap between different modalities and get the feature extractors for two heterogenous modalities.

Since VT-REID is also quite related to domain adaptation problems, they always adopt the structure where the all parameters are shared except for the last FC layers. For our cross-modal matching problem, we need to bridge the gap between two heterogenous modalities. Therefore, we assume that there exists some specific information in the shallow layers, so the parameters of shallow layers are specific for two different modalities. Then the multi-modality sharable feature representation is learnt with shared fully connected layers.

Hierarchical Cross-modality Metric Learning

Denote the features extracted from above two-stream network as $\{X_i\}_{i=1}^{K_1}$ for visible images and and $\{Z_j\}_{j=1}^{K_2}$ for thermal images, K_1 and K_2 represent the number of persons in each modality. And each person may have multiple images in each modality, thus each X_i and Z_j is represented by a $\mathcal{R}^{d \times n_{i(/j)}}$ matrix. To enhance the discriminability of the learned features and further improve the cross-modality matching accuracy, we further extend it with Hierarchical Cross-modality Metric Learning (HCML). The main idea about HCML is that the transform two heterogenous modalities (*modality-specific metrics*) into a consistent space that modality shared metric (*modality-shared metric*) can be subsequently learnt. Therefore, we formulate HCML as follows

$$\begin{aligned} F(V,T,M) =& \beta f(V,X) + \beta f(T,Z) + g(M,V,T,X,Z) \\ & s.t. \left\|v_i\right\|_2^2 \leq 1, \left\|t_i\right\|_2^2 \leq 1, M \succeq 0. \end{aligned} \tag{5}$$

where $V \in \mathcal{R}^{d \times N_d}$ and $T \in \mathcal{R}^{d \times N_d}$ represent the modality-specific transformation matrix for visible images and thermal images to handle the *intra-modality variations*, and transform the heterogenous features into a common space for modality-shared metric learning. N_d is a positive integer and usually defined by the dimension of the feature vector. v_i and t_i denote the i_{th} column vectors in V and T. M indicates the learnt modality-shared metric to handle *cross-modality discrepancy*. M is a positive semidefinite matrix as illustrated in (Liao and Li 2015) to make the learnt metric more robust. f and g separately represent the modality-specific and modality-shared terms, and β is the tradeoff between these two terms.

Modality-specific Terms. They aim to constrain the feature vectors of the same person within each modality should be compacted. Therefore, we formulate them by

$$f(V,X) = \frac{1}{N_x}\sum_{i=1}^{K_1}\sum_{k=1}^{n_i} h(\left\|V^T(x_{ik} - \bar{x}_i)\right\|_2^2) \tag{6}$$

$$f(T,Z) = \frac{1}{N_t}\sum_{j=1}^{K_2}\sum_{k=1}^{n_j} h(\left\|T^T(z_{jk} - \bar{z}_j)\right\|_2^2) \tag{7}$$

where $\bar{x}_i$ and $\bar{z}_j$ represent the mean vector of X_i and Z_j, respectively. N_x and N_t represent the total number of person images in each modality. n_i (n_j) is the number of person images of person i (j) in visible (thermal) modality. $h(x) = \frac{1}{\gamma}\log(1+e^{\gamma x})$ is a generalized logistic loss function, which is a smoothed approximation of the hinge loss function $[x]_+ = max(0,x)$ as illustrated in (Wang et al. 2017), γ is a sharpness parameter.

Modality-shared Term. It aims to learn the metric which could distinguish different persons from two heterogenous modalities after the transformation with modality-specific metrics. Therefore, we formulate it by

$$\begin{aligned} g(M,V,T,X,Z) = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\omega_{ij}h(y_{ij}(D_M(\bar{x}_i,\bar{z}_j) - \sigma)) \\ D_M(\bar{x}_i,\bar{z}_j) = (V^T\bar{x}_i - T^T\bar{z}_j)^T M (V^T\bar{x}_i - T^T\bar{z}_j) \end{aligned} \tag{8}$$

where y_{ij} is a binary indicator, $y_{ij} = 1$ if person i and j belong to the same person; otherwise, $y_{ij} = -1$. ω_{ij} is a weighting parameter to balance the asymmetric positive and negative sample pairs. Specifically, $\omega_{ij} = \frac{1}{N_{pos}}$ if $y_{ij} = 1$, and $\omega_{ij} = \frac{1}{N_{neg}}$ if $y_{ij} = -1$, and N_{pos} and N_{neg} are the total number of positive and negative samples pairs. σ is a constant positive bias, which is applied to ensure that D has a lower bound of zero.

Optimization of HCML

Since the optimization problem of (5) comprises of three different variables, it is hard to get its closed solution directly. Therefore, we solve it by an alternative iterative algorithm, which converts the problem into several sub problems where only one variable is involved. Detailed steps are as follows.

- **Initialize V, T and M.**

We initialize M as an identity matrix, while V and T are initialized in the same way. Alternatively, we could initialize V by solving the following problem.

$$\min_{V}\sum_{i=1}^{K_1}\sum_{k=1}^{n_i}\left\|V^T(x_{ik} - \bar{x}_i)\right\|_2^2 \quad s.t. V^TV = I \tag{9}$$

We solve it by an eigen-decomposition procedure. We construct the Lagrange function and let the derivation to zero, it becomes $Q_1V = \rho V$, and Q_1 is defined by

$$Q_v = \sum_{i=1}^{K_1}\sum_{k=1}^{n_i}(x_{ik} - \bar{x}_i)^T(x_{ik} - \bar{x}_i) \tag{10}$$

We select eigenvectors corresponding to the smallest N_d eigenvalues as V. Similarly, we get the initialization of T.

- **Fix V and T, update M.**

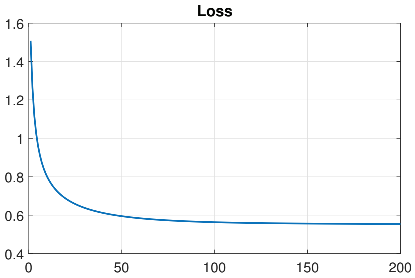

Figure 4: Convergence curve of HCML on RegDB dataset.

When V and T are fixed, the objective function regarding M can be written as

$$\min_{M} \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \omega_{ij} h(y_{ij}(D_M(\bar{x}_i,\bar{z}_j)-\sigma)) \quad s.t.M \succeq 0. \tag{11}$$

We derive an optimization algorithm based on accelerated proximal gradient (APG) algorithm (Beck and Teboulle 2009; Lan, Ma, and Yuen 2014) to solve problem (11). To accelerate the optimization procedure, an aggregation forward matrix S_t is introduced at each iteration t by

$$S_t = M_{t-1} + \frac{\xi_{t-1}-1}{\xi_{t-1}}(M_{t-1}-M_{t-2}) \tag{12}$$

where $\xi_t = (1+\sqrt{4\xi_{t-1}^2+1})/2$, and ξ_0 is initialized with zero (Beck and Teboulle 2009).

With a proper step size δ_t, a proximal operator is constructed by,

$$P_{\delta_t}(M,S_t) = g(S_t)+\langle M-S_t, \nabla g(S_t)\rangle + \frac{1}{2\delta_t}\left\|M-S_t\right\|_F^2 \tag{13}$$

where $\nabla g(S_t)$ is the gradient of $g(S_t)$. And $\langle A_1, A_2\rangle$ is the matrix inner product. $\|\cdot\|_F$ represents the Frobenius norm of a matrix. Based on the solution of S_t in (12), the distance metric M at the t-th iteration can be obtained by solving

$$\min_{M} P_{\delta_t}(M,S_t) \quad s.t.M \succeq 0 \tag{14}$$

Thus, we set $A_t = S_t - \delta_t \nabla g(S_t)$, and then conduct a singular value decomposition of $A_t = U_t \Lambda_t U_t^T$. The solution of (14) is

$$M_t = U_t \Lambda_t^+ U_t^T \tag{15}$$

where $\Lambda_t^+ = max\{0, \Lambda_t\}$ is defined to make sure that M_t is a positive semi-definite matrix in each iteration. The step size δ_t is initialized as 128 and adapted in a similar way as done in (Beck and Teboulle 2009).

- **Fix M and T, update V.**

With fixed M and T, we update V by using stochastic gradient descent (SGD) scheme. Samples are randomly selected at each iteration, and then V is updated by

$$V_{t+1} = V_t - l * (\frac{\partial f(V_t)}{\partial V_t} + \beta \frac{\partial g(V_t)}{\partial V_t}) \tag{16}$$

where l is the learning rate, and defined by $l = l_0 k^t$. And l_0, k are hyperparameters and t is the iteration number. In a similar way, we could update T with fixed V and M.

Convergence Analysis

The objective function of our HCML in problem (5) contains 3 different variables. If fix V and T, the convergence for M has been studied in (Beck and Teboulle 2009). When fix T and M, the function is also convex for V, and similar conclusion is also suitable for T. The objective function value is decreased at each iteration with a lower bound zero. Thus the convergence is ensured theoretically. To illustrate the convergence, we plot the objective function value at each iteration of our proposed method in Fig. 4. Empirically, we can see that the objective function value drops quickly and becomes stable after several iterations.

Experimental Results

Experimental Settings

Dataset. We use the publicly available RegDB dataset (Nguyen et al. 2017) for evaluation. RegDB is collected by dual camera systems, it contains 412 persons. For each person, 10 different visible light images are captured by a visible camera, and 10 different thermal images are obtained by a thermal camera. Thus, it simulates the practical applications that VT-REID suffers from the intra-modality variations and cross-view variations. Example images are shown in Fig. 2.

Settings. All the experiments are conducted following the standard evaluation protocol in existing visible image based re-id works, i.e., we randomly split the datasets into two halves, one for training and the other for testing. In the testing stage, the images from one modality were used as the gallery set while the ones from the other modality as the probe set. Since one person has multiple images in the gallery set, which means that one probe image may have multiple groundtruths. Therefore, we also adopt the mean average precision (mAP) for evaluation criteria besides the standard cumulated matching characteristics (CMC) curve. The procedure is repeated for 10 trials to achieve statistically stable results. Note that our setting is totally different from that in (Nguyen et al. 2017), since we focus on cross-modal re-identification.

Implementation details.[3] For two-stream CNN feature learning, we implement it on Tensorflow, we set the length of $fc2$ as 2048, which could achieve a slightly better performance than the original setting (4096). The trade-off parameter α of identity loss and contrastive loss it set as 0.2, and the maximum number of training epochs is set to 30. For efficiency considerations, we firstly reduce the dimension of features generated by $fc2$ from 2048 to 600 by PCA. For HCML, we set N_d as 600 to keep all the energies. Empirically, the balancing parameter β of modality-specific and modality-shared term is set to 0.2, since the major difficulty for cross-modal matching problem is to handle the cross-modality discrepancy. The smooth parameter γ of the logistic loss function is set to 1. And the learning rate to update V and T is initialized $l_0 = 0.1$, k is set to 0.9 in all our experiments. If without specification, we treat visible images as probe while thermal images as gallery set.

[3] Code is available on the first author's website.

Table 1: Evaluation of the proposed two-stream CNN network. The results are reported with re-identification rates (%) at rank r and mean average precision (mAP).

Methods	$r = 1$	$r = 10$	$r = 20$	mAP
HOG	13.49	33.22	43.66	10.31
MLBP	2.02	7.33	10.90	6.77
$fc1$	4.50	14.32	20.21	6.72
$fc2$	3.03	9.19	12.46	4.42
One-stream	13.11	32.98	42.51	14.02
Two streams	1.43	8.36	10.96	3.42
Concatenation	4.65	13.56	19.32	5.08
Substraction	1.60	7.62	14.32	3.21
Square	7.76	19.26	25.36	5.86
TONE_1	11.07	26.82	35.67	13.30
TONE_2	**16.87**	**34.03**	**44.10**	**14.92**

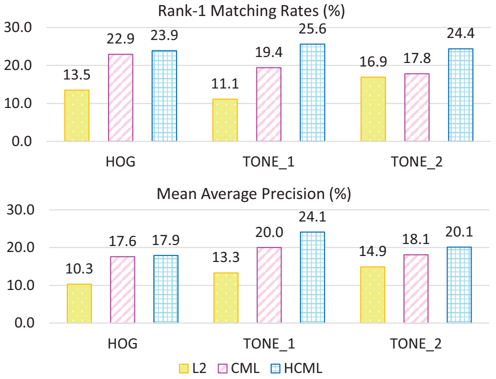

Figure 5: Evaluation of the HCML on the RegDB dataset based on rank-1 matching rates and mAP (%). HOG, TONE_1 and TONE_2 are selected for evaluation.

Evaluation of the TONE

This subsection evaluates the proposed TwO-stream CNN NEtwork (TONE) for the deep feature representation learning. Specifically, TONE_1 is extracted from $fc1$ layer and TONE_2 from $fc2$ layer in our network. Five baseline methods are adopted for comparison, including two shadow features and three baselines of deep features. 1) *HOG*[4]: As done in (Sarfraz and Stiefelhagen 2017), we finally get a 3986-dim vector for each image; 2) *MLBP*: Multi-scale LBP, each image is represented by a 7424-dim vector; 2) $fc1$ and $fc2$: Deep features pre-trained on ImageNet without fine-tuning, both $fc1$ and $fc2$ are included for comparison; 3) *One-stream*: Fine-tuning by simply using one stream network for both modalities; 4) *Two streams*: Fine-tuning two modalities with two separate streams separately. Additionally, other three two-stream CNN network structures with different configurations are also included, where the parameters of $fc2$ and $fc3$ are not shared. Similar to (Zheng, Zheng, and Yang 2016), we replace the contrastive layer with other choices, including two straight-forward ideas (*concatenation* and *substraction*) and *square* layer (Zheng, Zheng, and Yang 2016) with softmax cross-entropy loss. For fair comparison, we adopt Euclidean distance for comparison. The results are shown in Table 1.

As shown in Table 1, it can be seen that the re-identification performance can be improved by a large margin with the proposed two-stream CNN network. Compared to hand-craft features, deep features could capture more discriminative representations in an end-to-end feature learning manner. The rank-1 matching rate for our TONE is about 16.87%, and the mAP is about 14.92%. We suppose that our performance could be further improved with more training data, while the performance of HOG would be limited in scalable applications.

Compared to other baseline networks, the superiority of our TONE can attributed as two folds: 1) Partial shared layer parameters could help to learn multi-modality sharable feature representations (compared to *two streams*, where all weights are specific), 2) Partial specific layer parameters could well capture the discriminative intra-modality information in each specific modality (compared to *one-stream*, where all weights are shared). Additionally, compared to other two-stream network baselines (*Concatenation*, *Substraction* and *Square*) without shared fully connected layers, our performance is also superior as shown in Table 1. The main reason is that original architectures (Zheng, Zheng, and Yang 2016) designed for visible images based re-id are not suitable for our cross-modal re-identification problem.

Evaluation of the HCML

This subsection aims to show the effectiveness of the proposed matching model. Especially, to evaluate the modality-specific terms, which is quite important for re-identification task. Specifically, three kinds of features are included for comparison, which are TONE_1, TONE_2 and HOG. HCML represents the proposed hierarchical cross-modality metric learning method with both terms. CML means the metric learning without modality-specific term. The rank-1 matching rates and mAP are shown in Fig. 5.

Three main observations: (1) Compare with the Euclidean distance baselines, our proposed HCML could improve the performance often by a large margin. This can be attributed that the feature learning framework aims to model some local distributions with sample pairs, while the metric learning could learn a global discriminative matching space by balancing all the samples distributions. (2) Compared with CML, our HCML consistently improve the performance for all input features. The results validate that the proposed modality-specific terms could improve the performance for VT-REID. (3) Compared with shallow features, the overall performance of our proposed method is superior in most cases. Meanwhile, the gap between different features is also reduced with the metric learning step, which could verify the effectiveness of HCML further. Additionally, we find that TONE_1 is slightly better than that of TONE_2. This phenomena can be attributed to the overfitting problems suffered by CNN methods, the shallower layer features own better generalization ability.

[4] HOG and MLBP are extracted with the VLFeat toolbox.

Table 2: Comparison with other cross-modal matching methods. Re-identification rates (%) at rank r and mAP (%).

Methods	$r = 1$	$r = 10$	$r = 20$	mAP
L2	16.87	34.03	44.10	14.92
XQDA	21.94	45.05	55.73	**21.80**
MLAPG	17.82	40.29	49.73	18.03
GSM	17.28	34.47	45.26	15.06
SCDL	8.06	22. 09	28.89	10.03
rCDL	9.47	22. 96	29.42	10.26
HCML	**24.44**	**47.53**	**56.78**	20.80

Comparison with the state-of-the-arts

This subsection demonstrates the superior performance of the proposed cross-modal re-identification method. Some state-of-the-art cross-modal recognition methods are selected for comparison. As mentioned in Section , three metric learning methods (XQDA (Liao et al. 2015), MLAPG (Liao and Li 2015)) and GSM (Lin et al. 2017) are included. Two dictionary learning methods: semi-coupled dictionary learning method (SCDL) (Wang et al. 2012) and re-coupled dictionary learning (rCDL) (Huang and Frank Wang 2013) are adopted. We also try to implement the DeepMatch (Sarfraz and Stiefelhagen 2017), but the performance is too bad to be reported. For fair comparison, all the methods adopt our TONE_2 feature as input. All results are shown in Table 2.

The results shown in Table 2 illustrate that the proposed method outperforms other cross-modal matching models in most cases. Specifically, the rank-1 matching rates can achieve 24.4%, while the mAP values of ours is slightly lower than XQDA (20.80% *v.s.* 21.80%). The advantages of our proposed hierarchical matching method can be attributed as two folds: 1) Integrating the modality-specific term for VT-REID problem could reduce the intra-modality intra-person variations, which improves the VT-REID performance. 2) Rather than directly modality-shared metric learning, jointly optimization by firstly transforming two heterogenous modalities into a consistent space helps to address the cross-view variation problem, while other baselines mainly focus on addressing the cross-modality discrepancy.

Discussion

Impact of the contrastive loss. This subsection evaluates the weighting parameter α of the identity loss and contrastive loss. The results are shown in Fig. 6. We have also tried a larger α, but the feature learning algorithm could not get stable convergent results. When $\alpha = 0$, the proposed network is degenerated a common two-stream network with some shared and specific weights. Fig. 6 illustrates that integrating the contrastive loss with a suitable weight could boost the performance compared to identity loss only. Specifically, when $\alpha = 0.2$, the best performance can be achieved, which is about 3% improvement for the rank-1 matching rate. Similar conclusions can be drawn from the variation trend of mAP values.

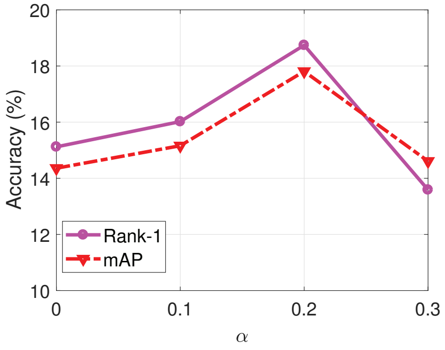

Figure 6: Influence of α for TONE on RegDB dataset.

Table 3: Evaluation of different query settings (%). (V-T: Visible to Thermal, T-V: Thermal to Visible.)

Settings	$r = 1$	$r = 10$	r=20	mAP
V-T (L2)	16.87	34.03	44.10	14.92
T-V (L2)	13.86	30.08	40.05	16.98
V-T (HCML)	24.44	47.53	56.78	20.08
T-V (HCML)	21.70	45.02	55.58	22.24

Insensitivity to query setting. Since our default setting is searching for thermal images with a visible image. We also evaluate the performance by changing the query setting, i.e., treating the thermal images as query set and visible images as gallery set. The rank-1 matching rates and mAP with our TONE_2 features are reported in Table 3. The results are close to the default settings (V-T module), with slightly lower rank-1 matching rates but higher mAP values. Therefore, we could see that the proposed method is robust to different query settings. The results illustrate the flexibility for practical heterogenous applications.

Modality fusion based re-identification. We also evaluate the performance for the modality fusion based re-identification with the learnt deep representations under similar settings in (Nguyen et al. 2017). We directly concatenate the features extracted from two modalities for re-identification. The rank-1 matching rates with our TONE_2 feature is 97.86% and mAP is 95.67%. With the proposed metric learning, the rank-1 matching rate could be improved to 98.72% while mAP is 98.23%. Results illustrate that the proposed method can also be applied to the modality fusion based re-identification. Compared to the L2-norm distance, HCML consistently improves the re-identification performance for both CMC and mAP values.

Conclusion

This paper investigates a cross-modal person re-identification problem, which is an important issue for many specific practical surveillance applications. Specifically, a two stage framework is proposed to address this issue. An improved two-stream CNN network aggregating the identity loss and contrastive loss is introduced to learn the multi-modality sharable feature representations. A hierarchical cross-modality discriminative metric learning method is presented to learn the matching model, which could simultaneously handle different variations. Extensive experiments validate the effectiveness of the proposed method compared with the baselines and state-of-the-arts.

Acknowledgement This work is partially supported by Hong Kong RGC General Research Fund HKBU (12202514), and National Natural Science Foundation of China (61562048).

References

Barbosa, I. B.; Cristani, M.; Del Bue, A.; Bazzani, L.; and Murino, V. 2012. Re-identification with rgb-d sensors. In *ECCVW*, 433–442.

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2(1):183–202.

Cao, Y.; Long, M.; Wang, J.; and Liu, S. 2017. Collective deep quantization for efficient cross-modal retrieval. In *AAAI*, 3974–3980.

Das, N.; Mandal, D.; and Biswas, S. 2017. Simultaneous semi-coupled dictionary learning for matching in canonical space. *IEEE Transactions on Image Processing (TIP)*.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2017. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, 2000–2006.

Huang, D.-A., and Frank Wang, Y.-C. 2013. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *ICCV*, 2496–2503.

Huo, J.; Gao, Y.; Shi, Y.; Yang, W.; and Yin, H. 2017. Heterogeneous face recognition by margin-based cross-modality metric learning. *IEEE TCYB*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Lan, X.; Ma, A. J.; Yuen, P. C.; and Chellappa, R. 2015. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. *IEEE TIP* 24(12):5826–5841.

Lan, X.; Ma, A. J.; and Yuen, P. C. 2014. Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. In *CVPR*, 1194–1201.

Lan, X.; Yuen, P. C.; and Chellappa, R. 2017. Robust mil-based feature template learning for object tracking. In *AAAI*, 4118–4125.

Lan, X.; Zhang, S.; and Yuen, P. C. 2016. Robust joint discriminative feature learning for visual tracking. In *IJCAI*, 3403–3410.

Li, S.; Xiao, T.; and et al. 2017. Person search with natural language description. In *CVPR*, 1345–1353.

Liao, S., and Li, S. Z. 2015. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 3685–3693.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2197–2206.

Lin, L.; Wang, G.; Zuo, W.; Feng, X.; and Zhang, L. 2017. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE TPAMI* 39(6):1089–1102.

Mogelmose, A.; Bahnsen, C.; Moeslund, T.; Clapés, A.; and Escalera, S. 2013. Tri-modal person re-identification with rgb, depth and thermal features. In *CVPRW*, 301–307.

Nguyen, D. T.; Hong, H. G.; Kim, K. W.; and Park, K. R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17(3):605.

Peng, C.; Gao, X.; Wang, N.; and Li, J. 2017. Graphical representation for heterogeneous face recognition. *IEEE TPAMI* 39(2):301–312.

Sarfraz, M. S., and Stiefelhagen, R. 2017. Deep perceptual mapping for cross-modal face recognition. *IJCV* 122(3):426–438.

Wang, S.; Zhang, L.; Y., L.; and Pan, Q. 2012. Semi-coupled dictionary learning with applications in image super-resolution and photo-sketch synthesis. In *CVPR*, 2216–2223.

Wang, Z.; Hu, R.; Liang, C.; Yu, Y.; Jiang, J.; Ye, M.; Chen, J.; and Leng, Q. 2016a. Zero-shot person re-identification via cross-view consistency. *IEEE Transactions on Multimedia* 18(2):260–272.

Wang, Z.; Hu, R.; Yu, Y.; Jiang, J.; Liang, C.; and Wang, J. 2016b. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *IJCAI*, 2669–2675.

Wang, Z.; Hu, R.; Chen, C.; Yu, Y.; Jiang, J.; Liang, C.; and Satoh, S. 2017. Person reidentification via discrepancy matrix and matrix metric. *IEEE Transactions on Cybernetics*.

Wu, A.; Zheng, W.-S.; and Lai, J.-H. 2017. Robust depth-based person re-identification. *IEEE TIP* 26(6):2588–2603.

Ye, M.; Liang, C.; Wang, Z.; Leng, Q.; Chen, J.; and Liu, J. 2015a. Specific person retrieval via incomplete text description. In *ICMR*, 547–550.

Ye, M.; Liang, C.; Wang, Z.; Leng, Q.; and Chen, J. 2015b. Ranking optimization for person re-identification via similarity and dissimilarity. In *ACM MM*, 1239–1242.

Ye, M.; Liang, C.; Yu, Y.; Wang, Z.; Leng, Q.; Xiao, C.; Chen, J.; and Hu, R. 2016. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia* 18(12):2553–2566.

Ye, M.; Ma, A. J.; Zheng, L.; Li, J.; and Yuen, P. C. 2017. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, 5142–5150.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, L.; Yang, Y.; and Tian, Q. 2017. SIFT meets CNN: A decade survey of instance retrieval. *IEEE TPAMI*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2016. A discriminatively learned cnn embedding for person re-identification. *arXiv preprint arXiv:1611.05666*.

Zhuang, Y.; Wang, Y.; Wu, F.; Zhang, Y.; and Lu, W. 2013. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, 1070–1076.