RGB-Infrared Cross-Modality Person Re-Identification

Ancong Wu\textsuperscript{1}, Wei-Shi Zheng\textsuperscript{2,5,6*}, Hong-Xing Yu\textsuperscript{2}, Shaogang Gong\textsuperscript{4}, and Jianhuang Lai\textsuperscript{2,3}

\textsuperscript{1}School of Electronics and Information Technology, Sun Yat-sen University, China
\textsuperscript{2}School of Data and Computer Science, Sun Yat-sen University, China
\textsuperscript{3}Guangdong Province Key Laboratory of Information Security, China
\textsuperscript{4}Queen Mary University of London, United Kingdom
\textsuperscript{5}Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
\textsuperscript{6}Collaborative Innovation Center of High Performance Computing, NUDT, China

\texttt{wuancong@mail2.sysu.edu.cn, wszheng@ieee.org, xKoven@gmail.com,}
\texttt{s.gong@qmul.ac.uk, stsljh@mail.sysu.edu.cn}

Abstract

Person re-identification (Re-ID) is an important problem in video surveillance, aiming to match pedestrian images across camera views. Currently, most works focus on RGB-based Re-ID. However, in some applications, RGB images are not suitable, e.g., in a dark environment or at night. Infrared (IR) imaging becomes necessary in many visual systems. To that end, matching RGB images with infrared images is required, which are heterogeneous with very different visual characteristics. For person Re-ID, this is a very challenging cross-modality problem that has not been studied so far. In this work, we address the RGB-IR cross-modality Re-ID problem and contribute a new multiple modality Re-ID dataset named SYSU-MM01, including RGB and IR images of 491 identities from 6 cameras, giving in total 287,628 RGB images and 15,792 IR images. To explore the RGB-IR Re-ID problem, we evaluate existing popular cross-domain models, including three commonly used neural network structures (one-stream, two-stream and asymmetric FC layer) and analyse the relation between them. We further propose deep zero-padding for training one-stream network towards automatically evolving domain-specific nodes in the network for cross-modality matching. Our experiments show that RGB-IR cross-modality matching is very challenging but still feasible using the proposed model with deep zero-padding, giving the best performance. Our dataset is available at \url{http://isee.sysu.edu.cn/project/RGBIRReID.htm}.

1. Introduction

Person re-identification (Re-ID) is an important field in video surveillance. A large number of models for Re-ID problem have been proposed, including feature learning\textsuperscript{[29, 48, 23]}, distance metric learning\textsuperscript{[55, 15, 22, 28, 23, 24]}, and end-to-end learning\textsuperscript{[20, 1, 47, 46]}. Most Re-ID methods are based on RGB-RGB matching, the most common single-modality Re-ID problem.

However, RGB-RGB Re-ID can be limited in surveillance when lighting is either poor or unavailable. For instance, RGB images become uninformative at night (Figure 1). In such a case, imaging devices without relying on visible light should be applied. Infrared (IR) cameras are commonly used in video surveillance systems. While depth images captured by RGB-D cameras such as Kinect are also independent of visible light, but they are rarely deployed because they are more expensive, used indoor only and with distance limitations. Since most surveillance cameras are able to automatically switch from RGB to IR mode in the dark, it is necessary to study RGB-IR cross-modality matching in 24-hour surveillance systems.

In this work, we introduce the RGB-IR cross-modality Re-ID problem. Although RGB-IR Re-ID is common and significant in real-world applications, to our best knowledge, it has been rarely explored and remains an open is-
sue. RGB-IR Re-ID is a very challenging problem due to the great differences between two modalities. Firstly, RGB and IR images are intrinsically distinct. See Figure 1, RGB images in the first row have three channels containing colour information of visible light, while IR images in the second row have one channel containing information of invisible light. Thus, they can be regarded as heterogeneous data. Secondly, from imaging principle aspect, the wavelength range of RGB and IR images is different. In existing Re-ID works, colour information is the most important appearance cue for identifying persons. However, in the RGB-IR Re-ID problem, this cue can hardly be used. As shown in Figure 1, even human can hardly recognise the persons by colour information. This leads to severe data misalignment within the same class. Moreover, viewpoint change, pose and exposure problems which cause large intra-class discrepancy in RGB-based Re-ID also bring difficulties to RGB-IR cross-modality Re-ID, resulting in a much more challenging problem. Although there exists a few Re-ID methods using IR images such as Jungling et al. [13]. They only consider the IR-IR video matching for Re-ID but does not consider the cross-modality RGB-IR Re-ID problem.

We first identify the challenge of RGB-IR Re-ID by conducting extensive evaluations on popularly used cross-modality methods. For this purpose, we have collected a new dataset called SYSU Multiple Modality Re-ID (SYSU-MM01) dataset. The comparison with existing commonly used Re-ID datasets is shown in Table 1. It contains 287,628 RGB images and 15,792 IR images of 491 persons captured in 6 cameras. To our best knowledge, this new RGB-IR Re-ID dataset provides for the first time a meaningful benchmark for the study of cross-modality RGB-IR Re-ID.

For cross-modality matching tasks, domain-specific modelling is important for extracting shared features for matching because of the domain shift. Considering using neural networks for cross-modality matching, we investigate and analyse the relation between different neural network structures, including two-stream structure and asymmetric FC layer structure, in which the domain-specific modelling exists but is designed manually. Alternatively, we propose a deep zero-padding method for training one-stream network tending towards evolving domain-specific structures automatically. Extensive experiments show the effectiveness of deep zero-padding, which outperforms the compared hand-crafted feature and deep models.

The contributions of this paper are: (1) We contribute for the first time a standard benchmark SYSU-MM01 for supporting the study of RGB-IR cross-modality Re-ID. We conducted extensive experiments to evaluate popular baseline deep learning architectures for cross-modality RGB-IR Re-ID. (2) We analyse three different network structures (one-stream structure, two-stream structure and asymmetric FC layer structure) and give insights on their effectiveness for RGB-IR Re-ID. (3) We propose deep zero-padding for evolving domain-specific structure automatically in one-stream network optimised for RGB-IR Re-ID tasks. Our experiments show that this approach for RGB-IR cross-modality Re-ID outperforms not only a standard one-stream network but also a two-stream network with explicit cross-domain learning and extra computational costs.

2. SYSU-MM01 Dataset

2.1. Dataset Description

SYSU-MM01 contains images captured by 6 cameras, including two IR cameras and four RGB ones. Different from RGB cameras, IR cameras work in dark scenarios. We show the details in Table 2 and some samples from each camera view in Figure 2. RGB images of camera 1 and camera 2 were captured in two bright indoor rooms (room 1 and room 2) by Kinect V1. For each person, there are at least 400 continuous RGB frames with different poses and viewpoints. IR images of camera 3 and camera 6 are captured by IR cameras in the dark. The IR images have only one channel, and they are different from 3-channel RGB images. Camera 3 is placed in room 2 in dark environment, while camera 6 is placed in an outdoor passage with background clutters. Camera 4 and 5 are RGB surveillance cameras placed in two outdoor scenes named gate and garden.

Observing the samples of the dataset, we can see clearly that the images of IR cameras (camera 3 and 6) are distinct from RGB images, in terms of both colour and exposure. Specifically, although camera 2 and 3 are in the same sce-
nario, the images of them suffer from dramatic colour shift and exposure difference. For example, the first person’s yellow clothes is distinct from her black trousers under the RGB camera, but this colour distinction is nearly eliminated under IR camera (Column 1, 2, Row 2,3 in Figure 2). Moreover, IR images have only one channel and might lose some texture details. The exposure of IR images captured at different distances is also an issue. These all introduce difficulties for the RGB-IR cross-modality Re-ID problem.

2.2. Evaluation Protocol

There are 491 valid IDs in SYSU-MM01 dataset. We have a fixed split using 296 identities for training, 99 for validation and 96 for testing. During training, all images of the 296 persons in training set in all cameras can be applied.

In the testing stage, samples from RGB cameras are for gallery set, and those from IR cameras are for probe set. We design two modes, all-search mode and indoor-search mode. For all-search mode, RGB cameras 1 and 2 are for gallery set and IR cameras 3 and 4 are for probe set. For outdoor-search mode, RGB cameras 2 and 3 (excluding outdoor cameras 4 and 5) are for gallery set and IR cameras 3 and 4 are for probe set.

For both modes, we adopt single-shot and multi-shot settings. For every identity under an RGB camera, we rank the images of the identity in descending order of similarities. For probe images of camera 3 skip the gallery images of camera 2. After computing similarities, we can get a ranking list according to descending order of similarities.

For indicating the performance, we use Cumulative Matching Characteristic (CMC) [32] and mean average precision (mAP). Notice that, for CMC under multi-shot setting, only the maximum similarity in all gallery images of the same person is taken to compute the rank list. We repeat the above evaluation 10 times with random split of gallery and probe set and compute the average performance finally.

3. Network Structure Comparison on Cross-Modality Modelling

We investigate deep learning network architectures for the task of RGB-IR cross-modality Re-ID. In particular, we examine three commonly adopted network structures for visual recognition and cross-modality learning. We further exploit the idea of deep zero-padding for model training and give insights on its impact on cross-modality matching task.

3.1. Common Deep Model Network Structures

In the past few years, a large number of deep models have been proposed for visual matching and cross-modality modelling, and have achieved satisfactory performance in many tasks. The most commonly used structures can mainly be categorized into 3 types. All structures that we are going to discuss are shown in Figure 3.

**One-stream Structure.** One-stream structure is the most commonly used in vision tasks. As shown in the first network in Figure 3, there is single input and all parameters are shared in the whole network. Representative networks include AlexNet [16], VGG [38], GoogleNet [40], ResNet [9] and so on, which perform well in classification, detection, tracking and many other tasks. In the field of Re-ID, JSTL-DGD [47], one of the state-of-the-art network, uses one-stream structure as well. Generally, in these tasks, the inputs to the network are RGB images, which are of the same modality. So sharing all parameters in the network is appropriate for these tasks.

**Two-stream Structure.** Two-stream structure is commonly used in cross-modality matching tasks. As shown in the second network in Figure 3, there are two inputs, corresponding to data in two different domains. In the shallower layers, the parameters of network are specific for each domain. In the deeper layers, shared parameters are used. The generalized similarity net [26] proposed by Lin et al. for cross-
domain visual matching including the Re-ID task is one of the representative structure of this type. Networks with two inputs similar to two-stream structure are also favorable in Re-ID tasks, for example, Ahmed’s net [1], SIR-CIR net [42], gated siamese net [41], etc. Note that except for Lin’s structure [26], most of them prefer sharing parameters in domain-specific layers. This is not exactly identical to our definition of two-stream structure. The reason may be, although the images are from different cameras, they are all of the same modality of RGB images. Compared to one-stream structure, two-stream structure achieves two things, domain adaptation and discriminative feature learning. It is assumed that the domain-specific network can extract shared features for different domains, and then the shared network can extract discriminative features for matching.

**Asymmetric FC Layer Structure.** Asymmetric FC layer model is also used in multi-domain tasks, for example, MDM-Net [33] for multi-domain tracking, CVDCA [2] for Re-ID and IDR [10] for VIS-NIR face recognition, etc. As shown in the third network in Figure 3, the structure shares nearly all parameters except for the last FC layer. This design assumes that the feature extraction for different domains can be the same and domain adaptation is achieved in feature level. This order of feature extraction and domain adaptation is different from two-stream structure.

### 3.2. Analysis of Network Structures

**– Connection of One-stream and Two-stream Structures in special case.** The three structures discussed above seem to be different, we find interestingly that all structures can be represented by one-stream structure in the forward propagation process when the following assumption is hold:

**Assumption 1.** A domain selection sub-network would exist somewhere in a network, which can automatically select samples of the corresponding domain as input, and the domain selection sub-network is fixed.

Under Assumption 1, we firstly give a simple example how one-stream network can perform as two-stream network in forward propagation. As shown in Figure 4, on the left is a simplified two-stream network: two fully connected networks, each with a specific layer (blue and red) and a shared layer (green). On the right is a one-stream network which can be conditionally equivalent to the two-stream one in forward propagation, in which there is a domain selection sub-network for selecting the following domain-specific structure. We first define some symbols for illustration. Let \( x_{d1} \in \mathbb{R}^d \) and \( x_{d2} \in \mathbb{R}^d \) denote the input of domain1 and domain2, respectively. We define a domain indicator \( y_{ind} \) as a vector with two elements, of which the value is \([1, 0]^T\) or \([0, 1]^T\) indicating domain1 or domain2, respectively. Let \( f_{sel}(x, y_{ind}) \) denote the domain selection sub-network, implementing the following function:

\[
 f_{sel}(x, y_{ind}) = \begin{cases} 
 [I_x, O_{d1}]^T x, & y_{ind} = [1, 0]^T \\
 [O_{dt}, I_{d2}]^T x, & y_{ind} = [0, 1]^T. 
\end{cases} \tag{1}
\]

The equation above suggests that if the domain selection sub-network is fixed, the two-stream network can be represented by one-stream network in forward propagation.

**– Analysis of One-stream Structure in General Case.** The assumption we hope above is less feasible. Now, we drop this assumption and analyse the domain-specific property of one-stream network. For cross-modality matching tasks, domain-specific modelling is important for extracting shared components for matching because of domain shift. Generally, in neural networks, e.g., two-stream and asymmetric FC layer structure, this is modelled by domain-specific structures. Thus we intend to analyse the domain-specific modelling in one-stream network. Our analysis is based on the following relaxed assumption:

**Assumption 2.** As shown in Figure 5, for a one-stream network dealing with inputs of two domains, we categorize the output nodes of each layer into three types, domain1-specific nodes, domain2-specific nodes and shared nodes. The categorization depends on whether the response of the node is domain-specific. Let \( x_{d1}^{(l)} \) and \( x_{d2}^{(l)} \) denote the input to layer \( l + 1 \) of domain1 and domain2, respectively. For example, \( x_{d1}^{(0)} \) and \( x_{d2}^{(0)} \) are inputs of the whole network. Let \( y_{ind}^{(l)} \) denote the \( i \)-th node in layer \( l \) and \( f_{out}(x_{ind}^{(0)}, i, l) \) denote the output of \( y_{ind}^{(l)} \) with the network input \( x_{ind}^{(0)} \), we have:

\[
 f_{out}(x_{ind}^{(0)}, i, l) = \sigma \left( \sum_j w_{j,i}^{(l-1)} f_{out}(x_{ind}^{(0)}, j, l-1) + b_i^{(l-1)} \right), \tag{2}
\]

where \( \sigma(\cdot) \) is the activation function, \( w_{j,i}^{(l-1)} \) and \( b_i^{(l-1)} \) are weight and bias parameters of layer \( l - 1 \). The type of node \( y_{ind}^{(l)} \) is defined by:

\[
 \text{type}(y_{ind}^{(l)}) = \begin{cases} 
 \text{domain1-specific}, & f_{out}(x_{d1}^{(0)}, i, l) \equiv 0 \\
 \text{domain2-specific}, & f_{out}(x_{d2}^{(0)}, i, l) \equiv 0 \\
 \text{shared}, & \text{otherwise}.
\end{cases} \tag{3}
\]

For domain1-specific nodes, we use identity sign in \( f_{out}(x_{d2}^{(0)}, i, l) \equiv 0 \), which means that for any input of domain2, the output of node \( y_{ind}^{(l)} \) is always zero.

Under Assumption 2, we define some symbols for analysis. Let \( L \) denote the loss function. Let \( o_i^{(l+1)} \) denote the
output of the $i$-th node before activation function in layer $l + 1$, $x^{(l)}$ denote the input to layer $l + 1$ and $w_i^{(l)}$ and $b_i^{(l)}$ denote the weight and bias parameters, i.e., $o_i^{(l+1)} = (w_i^{(l)})^T x^{(l)} + b_i^{(l)}$. Using the above defined categorization, without loss of generality, $x^{(l)}$ can be factorized into three parts $x^{(l)} = [x^{(l)\text{, }1\text{spec} }; x^{(l)\text{, }2\text{spec} }; x^{(l),s}]$ in which the three components denote the domain1-specific, domain2-specific and shared nodes, respectively. We can also denote $w_i^{(l)}$ as $w_i^{(l)} = [w_i^{(l),1\text{spec} }; w_i^{(l),2\text{spec} }; w_i^{(l),s}]$.

For an input of the network $x^{(0)}_{d1}$ in domain1, according to the categorization definition, $x^{(l)\text{, }2\text{spec}} = 0$ because for the output of each domain2-specific node, $f_{out}(x^{(0)}_{d2}, i, l) \equiv 0$. In the forward propagation process, the output of layer $l + 1$ is

$$o^{(l+1)}_i = (w_i^{(l),1\text{spec}})^T x^{(l)\text{, }1\text{spec}} + (w_i^{(l),s})^T x^{(l),s} + b_i^{(l)}.$$  

(4)

For an input of the network $x^{(0)}_{d2}$ in domain2, similarly, we have

$$o^{(l+1)}_i = (w_i^{(l),2\text{spec}})^T x^{(l)\text{, }2\text{spec}} + (w_i^{(l),s})^T x^{(l),s} + b_i^{(l)}.$$  

(5)

In the back propagation process, for input of the network $x^{(0)}_{d1}$ in domain1,

$$\frac{\partial L}{\partial w_i^{(l),1\text{spec}}} = \frac{\partial L}{\partial o_i^{(l+1)}} \frac{\partial o_i^{(l+1)}}{\partial w_i^{(l),1\text{spec}}} = \frac{\partial L}{\partial o_i^{(l+1)}} x_i^{(l)\text{, }1\text{spec}},$$  

(6)

$$\frac{\partial L}{\partial w_i^{(l),s}} = \frac{\partial L}{\partial o_i^{(l+1)}} \frac{\partial o_i^{(l+1)}}{\partial w_i^{(l),s}} = \frac{\partial L}{\partial o_i^{(l+1)}} x_i^{(l),s},$$  

(7)

$$\frac{\partial L}{\partial w_i^{(l),2\text{spec}}} = \frac{\partial L}{\partial o_i^{(l+1)}} \frac{\partial o_i^{(l+1)}}{\partial w_i^{(l),2\text{spec}}} = \frac{\partial L}{\partial o_i^{(l+1)}} x_i^{(l),2\text{spec}} = 0.$$  

(8)

From the analysis above, we have two conclusions: (1) In forward propagation, as shown in Figure 5, the weight parameters $w_i^{(l),1\text{spec}}$ (blue connections) and $w_i^{(l),2\text{spec}}$ (red connections) only have impact on input of corresponding domain, which is similar to the domain-specific parameters in two-stream networks. While for $w_i^{(l),s}$ (green connections), it has impact on both two domains, which is similar to the shared parameters in two-stream networks. Thus, the network can implicitly control the domain-specific structure by domain-specific nodes and control the shared structure by shared nodes. (2) In backward propagation, if a node is domain2-specific, with input in domain1, its corresponding weight parameters will not be updated because the gradient is zero. That means the training samples of the other domain would not influence the implicit domain-specific structure. Note that for an input $x^{(0)}_{d2}$, the same conclusion can be drawn in a similar way.

**Remark 1.** A one-stream network may implicitly learn and evolve the domain-specific and shared structures in the network if the three types of nodes defined by Equation (3) are assumed to be existed in the network.

**Remark 2.** Considering two-stream structure and asymmetric FC layer structure, they are designed manually and fixed during training. Moreover, the domain-specific structure of two domains are decoupled, while the shared structure is completely identical. In contrast, if one-stream structure can implicitly learn the structure, the implicit structures corresponding to different domains are partially coupled by shared nodes and shared bias parameters (Equations (4) and (5)), which can provide more flexibility in training for cross-modality matching tasks.

### 4. Deep Zero-Padding

#### 4.1. Analysis of Zero-Padding as Network Input

Since the node type we define in the last section (Equation (3)) is very optimal based on the assumption that $f_{out}(x^{(0)}_{d1}, i, l) \equiv 0$ or $f_{out}(x^{(0)}_{d2}, i, l) \equiv 0$, and how to make the network learn such nodes with the domain-specific property in training stage remains an important problem. In most cases, one-stream network is applied in single-domain tasks, which treats all samples equally so that generally domain-specific nodes may not be learned.

As analyzed in the previous sections, the structures of two-stream network and asymmetric FC layer network are designed manually and fixed during training, while one-stream network can evolve the network structure implicitly by learning domain-specific nodes, which may generate more optimal structure. For this purpose, we propose to use zero-padding input to stimulate the domain-specific response. As shown in Figure 5, for inputs from two domains $x_{d1} \in \mathbb{R}^d$ and $x_{d2} \in \mathbb{R}^d$, we apply zero-padding as follows:

$$x_{d1}^{pad} = [x_{d1}^T, O_1 \times d]^T, x_{d2}^{pad} = [O_1 \times d, x_{d2}^T]^T.$$  

(9)

If we regard the network input as a prior-layer (or called the 0-th layer), then all the nodes in such a prior-layer will be definitely categorized as domain-specific nodes according to our definition in Equation (3). Now, what is the case for the nodes on the next layer? Indeed, it is hard to mathematically tell what it is, but we find that with the zero-padding as network input, the nodes in the networks are
more possibly becoming domain-specific nodes. Here we continue the analysis in Section 3.2, after applying activation function \( \sigma(\cdot) \) to Equations (4) and (5) we have:

\[
x_{d1,i}^{l+1} = \sigma((w_i^{(l)})_{1spe}x_{d1}^{l} + (w_i^{(l)})_{2spe}x_{d2}^{l} + b_i^{(l)}),
\]

\[
x_{d2,i}^{l+1} = \sigma((w_i^{(l)})_{2spe}T(x_{d2}^{l}),s + (w_i^{(l)})_{1spe}T(x_{d1}^{l}),s + b_i^{(l)}).
\]

(10)

where \( x_{d1,i}^{l+1} \) and \( x_{d2,i}^{l+1} \) denote the output of the \( i \)-th node in layer \( l+1 \) with input from domain1 and domain2.

We find that for \( x_{d1,i}^{l+1} \), there is an independent projection \( (w_i^{(l)})_{1spe}T(x_{d1}^{l}),s \) which has no overlap with the part inside the activation function for \( x_{d2,i}^{l+1} \). This means \( w_i^{(l)} \), becomes a free variable independent of the terms in the activation function for \( x_{d2,i}^{l+1} \). Thus it can provide more flexibility for neural networks to make \( f_{out}(x_{d1}^{(0)}, i, l + 1) > 0 \) and \( f_{out}(x_{d2}^{(0)}, i, l + 1) \equiv 0 \) to meet the requirement for domain1-specific nodes as compared to Equation (3). It would be easier for neural network to spread the domain specific-nodes in deeper layers. Since the zero-padding not only appears in the input, but also can spread in the network. We call this method deep zero-padding.

Of course, our deep zero-padding only makes neural networks more flexible in spreading domain-specific nodes in the network but not force to. Actually, our neural networks learning empirically support this. As shown in Figure 7 and Figure 8, deep zero-padding helps the network learn domain-specific nodes more easily than without zero-padding. The details will be illustrated later in Section 4.2.

4.2. Deep Zero-Padding for RGB-IR Re-ID

In our task RGB-IR cross-modality Re-ID, a channel is corresponding to a node in FC layer for convolutional neural network. For images, zero-padding is conducted in channel level. As illustrated in Figure 6, RGB image is converted to gray-scale image and placed in the first channel, and then zero-padding image is placed in the second channel. For infrared image, it is placed in the second channel and zero-padding image is placed in the first channel.

To show the effectiveness of deep zero-padding, we visualize the feature map of ResNet-6 in our experiments and compare the differences between deep zero-padding and original single-channel input. In Figure 7, we compute the average feature maps of 50 different persons on our dataset, and show all the 16 feature maps of the first and second convolution layers. As defined in Equation (3), we can categorize the domain-specific channels indicated by the bounding boxes. It is evident that, deep zero-padding helps to learn more domain-specific channels than single-channel input.

To quantify the domain-specific nodes in the network, we calculate the proportion of domain-specific nodes in each layer. Both a small (strict) threshold and a large (loose) threshold were set to determine whether the node is domain-specific according to Equation (3). The relation between the proportion of domain-specific nodes and layer depth is shown in Figure 8. It can be observed that, the domain-specific nodes mainly appear in shallower layers. It is reasonable that the network prefers shared structure after layer 6. Using deep zero-padding helps to generate more domain-specific nodes, while the proportions without zero-padding are low in most layers. The details about the respective proportions of two domains are provided in supplementary.

Domain-specific nodes enable the network to convolve image from different domains using different filters, so as to better alleviate the differences (e.g., gradient orientations and exposure differences) in Figure 1 between two domains. The analysis above and experiment results in Table 3 show that, the network can learn domain-specific nodes easier with deep zero-padding and achieve better performance.

4.3. Comparison of Cross-Modality Learning

While cross-modality matching task has not drawn much attention in Re-ID problem, it has been studied a lot in other fields like information retrieval and face verification. Cross-modality retrieval (e.g. text-image, tag-image) plays an important role in information retrieval. The models for cross-modality retrieval can be classified into real-value representation and binary representation learning [43]. The former
one includes subspace learning methods [25, 30, 60] and deep learning frameworks [45, 6, 14, 12], while the latter one includes linear models [39, 36, 59, 51] and non-linear models [27, 50, 31]. Generalised similarity measure proposed by Lin et al. [26] is for cross-domain visual matching tasks, including RGB-RGB Re-ID task. Matching visual face versus near infrared ones (VIS-NIR) [17, 58, 10] is rather related to RGB-IR cross-modality Re-ID. Compared with VIS-NIR face recognition, RGB-IR Re-ID is much more challenging, due to lack of important colour information. The remaining useful cues may be body shape, which differs greatly with different viewpoints and poses.

In some multi-domain learning methods, e.g., HFA [18], CRAFT [3], zero-padding in feature level is applied and proved to be effective. These works are closely related to our deep zero-padding. In comparison, our zero-padding is done in raw image level and the domain-specific and shared learning are done by deep neural network.

5. Experiments

We conducted extensive evaluations of existing Re-ID and cross-domain matching models as baselines on our SYSU-MM01 dataset. Then, we evaluated and analysed the effectiveness of deep models, including the proposed deep zero-padding and three network structures discussed in Section 3. See Section 2.2 for detailed evaluation protocol.

5.1. Compared Models

**Baseline Models.** We evaluated three favorable handcrafted features and cross-domain metric learning models as baselines. Handcrafted features included HOG [4], LOMO [23] and HIPHOP [3]. Metric learning methods included KISSME [15], LFDA [34] and cross-domain models CCA [35], CDFE [25], GMA [37], SCM [51] and CRAFT [3].

**Deep Models.** We evaluated four deep models shown in Figure 3, including one-steam network, two-stream network, asymmetric FC layer network and the proposed deep zero-padding method (network structure is the same as one-stream network). We applied residual block in ResNet [9] as the base convolution block for all the four structures. The number of filters for each block is 16, 16, 64, 128, 256 and 512, respectively. The next layer is an FC layer of 256 dimensions used as feature. For these four networks, the loss function was softmax loss as in ResNet [9], which is commonly used and relatively stable. All of the hyper parameters were kept the same. As for the input of the first three networks, the images were converted to single-channel gray-scale image and adjusted the size to 224 × 224. For our proposed method, the input was further processed by deep zero-padding as introduced in Section 3.

5.2. Model Comparisons and Analysis

**Deep Models vs. Baseline Models.** We show comparative results in Table 3, including the rank-1, 10, 20 accuracies of CMC [32] and mean average precision (mAP).

From Table 3, we can see clearly that deep models outperformed baseline models with large margins. Specifically, the proposed deep zero-padding outperformed all the baselines by nearly 10% in terms of rank-1 accuracies.

All baseline models, i.e., handcrafted features with cross-domain metric learning methods, performed poorly: even the rank-1 accuracy of the best case failed to reach 10%. Note that LOMO feature contains rich colour information, and it performs very well in RGB-RGB Re-ID problem. Hence the results indicate that in RGB-IR matching, due to different imaging principles, the colour’s discrimination degrades largely. Although body shape and clothes textures can be used for identifying persons, low-level features are not discriminative enough for the RGB-IR cross-modality person Re-ID problem. As for the deep models, however, the best rank-1 accuracy can achieve 20.58% for indoor-search. Although the problem is challenging, deep models are feasible to deal with it.

**Deep Zero-Padding vs. Other Deep Strategies.** Although deep models outperformed baseline models on the whole, there were gaps among their performances to some extent.

In Table 3 we can see that the deep zero-padding out-performed two-stream network and asymmetric FC layer structure. Taking rank-1 accuracy in all-search mode under single-shot setting for instance, the gaps between deep zero-padding and two-stream/asymmetric FC layer are 3.15%/5.50%. It is possibly because two-stream and asymmetric FC layer structure require careful architecture design for this task, so their structures may not be optimal.

While with the same one-stream ResNet-6 network structure, deep zero-padding also showed superiority over the original one. The differences mainly took place in the training stage. Deep zero-padding generated two domain-specific channels in the input layer, helping the network evolve the domain-specific structure implicitly during the training stage. Since the domain-specific and shared struc-
Table 3. Performance under all-search and indoor-search mode. r1, r10, r20 denote rank-1, 10, 20 accuracies (%).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>All-search</th>
<th>Single-shot</th>
<th>Multi-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r1</td>
<td>r10</td>
<td>r20</td>
<td>r1</td>
</tr>
<tr>
<td>One-stream network (deep zero-padding)</td>
<td>Euclidean</td>
<td>14.80</td>
<td>54.12</td>
<td>71.33</td>
</tr>
<tr>
<td>One-stream network (deep zero-padding)</td>
<td>GSM</td>
<td>12.04</td>
<td>49.68</td>
<td>66.74</td>
</tr>
<tr>
<td>Two-stream network (2 specific + 4 shared blocks)</td>
<td>Euclidean</td>
<td>11.65</td>
<td>47.99</td>
<td>65.50</td>
</tr>
<tr>
<td>Asymmetric FC layer network</td>
<td>Euclidean</td>
<td>9.30</td>
<td>42.46</td>
<td>60.68</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>HIPHOP</td>
<td>5.20</td>
<td>23.11</td>
<td>32.95</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>CRAFT</td>
<td>1.80</td>
<td>14.56</td>
<td>26.29</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>KISSME</td>
<td>2.12</td>
<td>16.21</td>
<td>29.13</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>LPDA</td>
<td>2.33</td>
<td>18.35</td>
<td>35.38</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>CCA</td>
<td>2.74</td>
<td>18.91</td>
<td>32.51</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>CDFE</td>
<td>2.09</td>
<td>16.68</td>
<td>30.51</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>GMA</td>
<td>1.07</td>
<td>10.42</td>
<td>20.91</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>NCM</td>
<td>1.86</td>
<td>19.16</td>
<td>28.22</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>CRAN</td>
<td>2.59</td>
<td>17.93</td>
<td>31.50</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>LMO</td>
<td>1.75</td>
<td>14.14</td>
<td>26.63</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>HIPHOP</td>
<td>2.23</td>
<td>18.95</td>
<td>29.67</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>CRAFT</td>
<td>2.98</td>
<td>21.11</td>
<td>33.36</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>CDFE</td>
<td>3.64</td>
<td>23.18</td>
<td>37.28</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>GMA</td>
<td>1.04</td>
<td>10.45</td>
<td>20.81</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>NCM</td>
<td>2.24</td>
<td>18.12</td>
<td>26.07</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>CRAN</td>
<td>2.34</td>
<td>18.70</td>
<td>32.95</td>
</tr>
</tbody>
</table>

Table 4. Comparison of deep zero-padding and similar networks. r1 and r10 denote rank-1 and 10 accuracies (%).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Metric</th>
<th>Single-shot</th>
<th>Multi-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r1</td>
<td>r10</td>
<td>r20</td>
</tr>
<tr>
<td>One-stream network (deep zero-padding)</td>
<td>Euclidean</td>
<td>12.04</td>
<td>49.68</td>
</tr>
<tr>
<td>One-stream network (deep zero-padding)</td>
<td>CRAFT</td>
<td>11.79</td>
<td>50.86</td>
</tr>
<tr>
<td>One-stream network (deep zero-padding)</td>
<td>Euclidean</td>
<td>14.80</td>
<td>54.12</td>
</tr>
</tbody>
</table>

Deep Zero-Padding vs. Domain Indicator. As illustrated in Section 3, one-stream network may be able to work as two-stream network with existence of domain indicator. So we padded two additional channels to the input images as domain indicators. For an RGB image, the first channel was padded with all pixels equal to 255, and the second with 0, while for an IR image the first channel was padded with 0 and the second was padded with 255. This padding procedure explicitly provided domain indicators to the network. The performance comparison is shown in Table 4.

Table 4 shows that our proposed deep zero-padding method achieved the best performance. Input with domain indicator only achieved comparable performance as original one-stream network. This result indicates that deep zero-padding can exploit domain information more effectively.

6. Summary

To our best knowledge, this work is the first to identify the RGB-IR cross-modality Re-ID problem and introduce a new multi-modality Re-ID dataset named SYSU-MM01. The great difference between RGB and IR images makes RGB-IR cross-modality Re-ID formed as a very challenging problem. We have discussed and evaluated three common network structures for cross-domain tasks including one-stream structure, two-stream structure and asymmetric FC layer structure. We have analysed the connection between one-stream and two-stream structure and found that one-stream network can learn and evolve domain-specific structure implicitly if there exist domain-specific and shared nodes. We propose deep zero-padding to help one-stream network more likely evolve domain-specific nodes automatically so as to make the implicit network structure more suitable for the task, and this is an alternative and flexible way for cross-modality modelling as compared to the manually designed fixed structure of two-stream network. The experiments have shown that the one-stream network trained by deep zero-padding achieved the best performance.

Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC (61522115, 61472456, 61573387, 61661130157, U1611461), the Royal Society Newton Advanced Fellowship (NA150459), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157). The principal investigator for this work is Wei-Shi Zheng.
References


