IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 28, 2019, accepted April 30, 2019, date of publication May 3, 2019, date of current version May 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914670

Person Re-Identification Between Visible and Thermal Camera Images Based on Deep Residual CNN Using Single Input

JIN KYU KANG, TOAN MINH HOANG, AND KANG RYOUNG PARK
Division of Electronics and Electrical Engineering, Dongguk University, Seoul 100-715, South Korea

Corresponding author: Kang Ryoung Park (parkgr@dongguk.edu)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041921), in part by the Basic Science Research Program through NRF funded by the Ministry of Education (NRF-2017R1D1A1B03028417), and in part by the NRF Grant funded by the Korea Government (Ministry of Science and ICT) (NRF-2017R1C1B5074062).

ABSTRACT In recent years, numerous studies have been undertaken regarding person re-identification (ReID), an important issue for intelligent surveillance systems. Person ReID, however, is an extremely difficult problem because of variables such as different viewpoints and poses, and varying lighting in person regions in images that have been captured from remote distances. A majority of the studies have been performed for visible-light camera-based person ReID, which can be used only in a limited environment owing to the characteristics of a visible-light camera that are considerably dependent on the illumination. To overcome this problem, studies have been conducted for multimodal camera-based person ReID. However, because the previous studies used two or more input images, the computational complexity was high. This paper proposes a novel person ReID method that simplifies the convolutional neural network (CNN) structure by combining visible-light and thermal images as a single input. This method overcomes the limitation of visible-light camera-based person ReID using both a visible-light and thermal camera. To verify the performance of the proposed method, two open databases, the DBPerson-Recog-DB1, and Sun Yat-sen University multiple modality Re-ID (SYSU-MM01) databases were used. The method proposed in this study demonstrated excellent performance compared to the conventional methods.

INDEX TERMS Person re-identification (ReID), CNN, multimodal camera (RGB-IR).

I. INTRODUCTION

Paralleling the advancement in intelligent surveillance systems, different studies have been undertaken regarding person re-identification (ReID). There are numerous studies [1], [2], [4]–[11] that have demonstrated robust performance for varying viewpoints and poses based on visible-light cameras; however, they experience the problem of being vulnerable at night without sufficient light or in outdoor environments where the illumination changes. Furthermore, in the case of a person changing clothing, a color-sensitive visible-light camera has difficulty with person ReID. To address these problems, studies have been actively performed on person ReID using multimodal cameras [12]–[14]. A multimodal camera is a device used to overcome the limitation of a visible-light camera by adding a depth camera or thermal camera to a visible-light camera. In the case of multimodal cameras, studies have been performed on person ReID for images acquired from the same view and viewpoint from different cameras [12], [14], and for person ReID between two or more types of cameras installed at different places and views [13], [23]. In the case of the former, the person ReID is relatively easy because there is virtually no difference in the viewpoint, pose, and illumination, whereas in the latter case, there are numerous difficulties in the person ReID because the view, pose, and illumination, as well as the characteristics of the image data are different. The structure of this paper is organized as follows. In Section II, different conventional studies on person ReID are analyzed in further detail. In Section III, the contribution of this study is presented and in Section IV, the proposed method is described in detail. In Section V, the experimental

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

results with analyses are discussed, and lastly, in Section VI, a conclusion is provided.

II. RELATED WORKS

As a study in the early stage, Huang and Russel employed the Bayesian method to track the same object from two camera views [1]. Later, Zajdel et al. designed the probabilistic relationship between a person and features (color and spatial-temporal features) for person ReID using a dynamic Bayesian network model [2]. Deep learning in image classification demonstrated a high performance in the 2012 ImageNet Large Scale Visual Recognition Competition (ILSVRC) [3] and influenced the studies on person ReID. Yi et al. used a Siamese neural network to determine if a pair of input images received from two cameras was indicating the same person [5]. In a similar fashion, studies have been performed diversely on person ReID starting from multi-camera tracking studies to deep learning-based studies [6].

In recent years, the majority of person ReID studies have proposed deep models that are robust to varying poses and viewpoints based on visible-light-cameras. Ahmed et al. proposed a binary classification model that expresses two feature maps extracted from two different images as a concatenated feature map, and determines if the persons are the same or different persons [7]. Varior et al. proposed a binary classification model that uses the discriminative capability of the local features based on the contextual information of the input images using a long short-term memory (LSTM) architecture [8]. Wang et al. proposed a triplet loss-based method that increases the inter-class variation and decreases the intra-class variation by dividing an image set into query, positive, and negative sets [29]. Cheng et al. proposed a person ReID method based on an improved triplet loss function based on multi-channel parts composed of a full-body channel and local body-part channel [10]. Chen et al. proposed a quadruplet loss-based person ReID that can increase the inter-class variation and decrease the intra-class variation by adding an extra negative pair to the conventional triplet loss [11].

As described above, in the majority of cases, the person ReID methods are typically based on a multi-input or multi-stream. Here, multi-input and single-input refer to the cases of using two or more images and one image, respectively, as CNN input. Furthermore, multi-stream has a structure whereby two or more networks start with independent structures and through feature fusion in a convolutional layer or fully-connected layer, operate as a single network (e.g., Siamese network) Single-stream has a structure that operates as a basic single network (e.g., AlexNet, VGG, GoogleNet). Unlike the pervious methods, Chen et al. proposed a joint representation learning method that facilitates binary classification with a single image by linking two input images in a horizontal direction [9]. As mentioned above, different methods exist to solve the person ReID problem based on visible-light cameras. However, they are limited by environmental conditions because of the characteristic of visible-light-camera being sensitive to illumination changes. For example, a visible-light camera cannot function properly at night when there is no light. This problem renders the value of intelligent surveillance systems useless at night, when crime risk is typically high. There are multimodal camera-based person ReID studies that use a thermal or depth camera together to solve this problem [12]–[14], [23]. The multimodal cameras used for Person ReID include visible-depth module, visible-thermal module, and visible-depth-thermal module. Wu et al. proposed a method of improving the performance of person ReID using an invariant body shape and skeleton information from a depth camera to solve the vulnerability of visible cameras in extreme illumination or changed clothing [12]. Ye et al. proposed a dual-path network that facilitates person ReID between visible and thermal images with ranking loss and identity loss by extracting features from the visible and thermal images, respectively [13]. Møgelmose et al. proposed a tri-model-based person ReID that uses all of the visible, depth, and thermal features [30]. Tri-model is a system that performs person ReID by fusing the features extracted using a histogram, Mean Shift [32], [38], and speeded-up robust features (SURF) [31] from visible, depth, and thermal images. Similarly, there are different studies that use multimodal cameras to overcome the limitations of the visible-light camera. However, in the case of a depth camera, the cost is significant and because of the limitation of distance measuring, it is difficult to use in an outdoor remote distance environment. Therefore, for intelligent surveillance systems, ReID studies have been conducted with visible-light and thermal cameras. Multimodal camera-based person ReID methods are also typically based on multi-input and multi-stream. However, Wu et al. proposed a deep zero-padding method that composes gray-scaled visible and thermal images into one 2-channel image pair [23]. This algorithm implements specific nodes of the visible-light and thermal images using a zero vector, thereby obtaining similar functions as the multi-stream structure and reducing the size of the network. The proposed method also focuses on a one-stream structure for simplifying the network. Moreover, to overcome the limitation of person ReID between visible-light images, this paper proposes ReID between a visible-light and thermal image method. To reduce the computational complexity, this method uses a one-stream network that has an inter-channel pair between the visible-light and thermal images (IPVT-1), an intra-channel pair between the visible-light and thermal images (IPVT-2), and an inter-channel pair and intra-channel pair between the visible-light and thermal images (IIPVT), as inputs. Table 1 summarizes the advantages and disadvantages of the methods proposed in the conventional studies and this study for person ReID.

III. CONTRIBUTIONS

Our research is novel in the following four aspects compared to previous works.

- We propose person ReID method based on a single-input and one-stream CNN that generates an IPVT-1, IPVT-2, and IIPVT from visible-light and thermal images, and

TABLE 1. Comparison of proposed and previous research on person ReID.

Category			Advantage	Disadvantage
Visible camera-based	Multi-input & multi-stream	Siamese LSTM [8]	Using Siamese LSTM, which can leverage the contextual information to enhance the discriminative capability of the local features	- Weak for extreme lighting conditions - Dependent on the colors of clothing, hair, and such.
		Multi-channel with improved triplet loss [10]	By learning both entire and part of human body, the conventional triplet loss method is improved	
	Multi-input & one-stream	Quadruplet loss [11]	By adding a negative-negative pair in the conventional triplet loss, the intra-class distance is reduced and the inter-class distance is increased	
	Single-input & one-stream	Joint representation learning [9]	One stream network by raw image rather than handcrafted features	
Multimodal camera-based	Multi-input & handcrafted feature	Visible-depth images [12]	Using depth information is less vulnerable to extreme illumination and changed clothing	Depth camera is difficult to use outdoors at a distance because of limitation of infrared sensor and low resolution image
	Multi-input & multi-stream	Dual-constrained top-ranking [13]	A model that facilitates end-to-end learning based on the visible-light and thermal cameras	High computational complexity compared to one stream structure because of dual path network
	Single-input & one-stream	Deep zero padding [23]	One stream structure is facilitated by applying the deep zero padding method	Lack experiments on open databases
		Proposed method	- Robust to illumination because of a feature learning method matching the visible-light and thermal cameras - Low computational complexity - Robust performance by considering both the inter-channel and intra-channel features	With Sun Yat-sen University multiple modality Re-ID (SYSU-MM01) database, accuracy improvement was required for the rank 10 and rank 20

uses them as a single input. Its computational complexity is low compared to the multimodal camera-based multi-stream methods.

- An IPVT-1 is an inter-channel pair composed of visible-light and thermal images, and an IPVT-2 is an intra-channel pair. By combining an IPVT-1 and IPVT-2, the two pair types are composed into a single IIPVT. Using an IPVT-1, IPVT-2, and IIPVT, the distance of the positive and negative pairs is increased, thereby improving the performance of the person ReID.
- The IIPVT, IPVT-1, and IPVT-2 are comparatively evaluated to determine the optimized image pair for person ReID by one-stream CNN. Moreover, the accuracy of the person ReID is improved using multi-scale Retinex (MSR)-filtered input images and inputs of deep CNN.
- For a fair comparative evaluation with other research, the labeled information of the IPVT-1, IPVT-2, and IIPVT and the trained CNN model were opened to other researchers, as presented in [33].

IV. PROPOSED METHOD

A. OVERALL PROCEDURE OF PROPOSED SYSTEM

Fig. 1 displays the overall procedure of the system proposed in this paper. For input, this system uses a person region detected in images that have been captured at different places and times or captured simultaneously from two or more cameras consisting of a visible-light and thermal camera. Whereas the most important visible-light image information is the color information of a person, a thermal image provides different information, namely, the thermal energy of the person and thermal energy of the surrounding environment. To extract correlation features from mutually different image features such as this, the IPVT-1, IPVT-2, and IIPVT were designed. If the learning is performed using the IPVT-1, IPVT-2, or IIPVT as the input image for the CNN, it is possible to extract common features of different images. From the features extracted in this manner, similarity and dissimilarity factors can be obtained by viewing the softmax layer. If the IPVT-1, IPVT-2, or IIPVT for the same person is used as an input image, the similarity factor of the heterogeneous feature increases. Conversely, if the IPVT-1, IPVT-2, or IIPVT for a different person is used as the input image, the dissimilarity factor of the heterogeneous feature increases. That is, CNN learning is facilitated in the direction of decreasing the intra-class variation and increasing the inter-class variation. From this process, a person image captured in a thermal image can be found for the same person in a person image captured in visible-light.

B. IPVT-1, IPVT-2, AND IIPVT FOR HETEROGENEOUS FEATURE LEARNING

Prior to feature learning between the visible-light and thermal images, a problem regarding how to solve the difference of the characteristics possessed by the visible-light and thermal data must be resolved. As described earlier, the visible-light data express red, green, and blue-based color information,

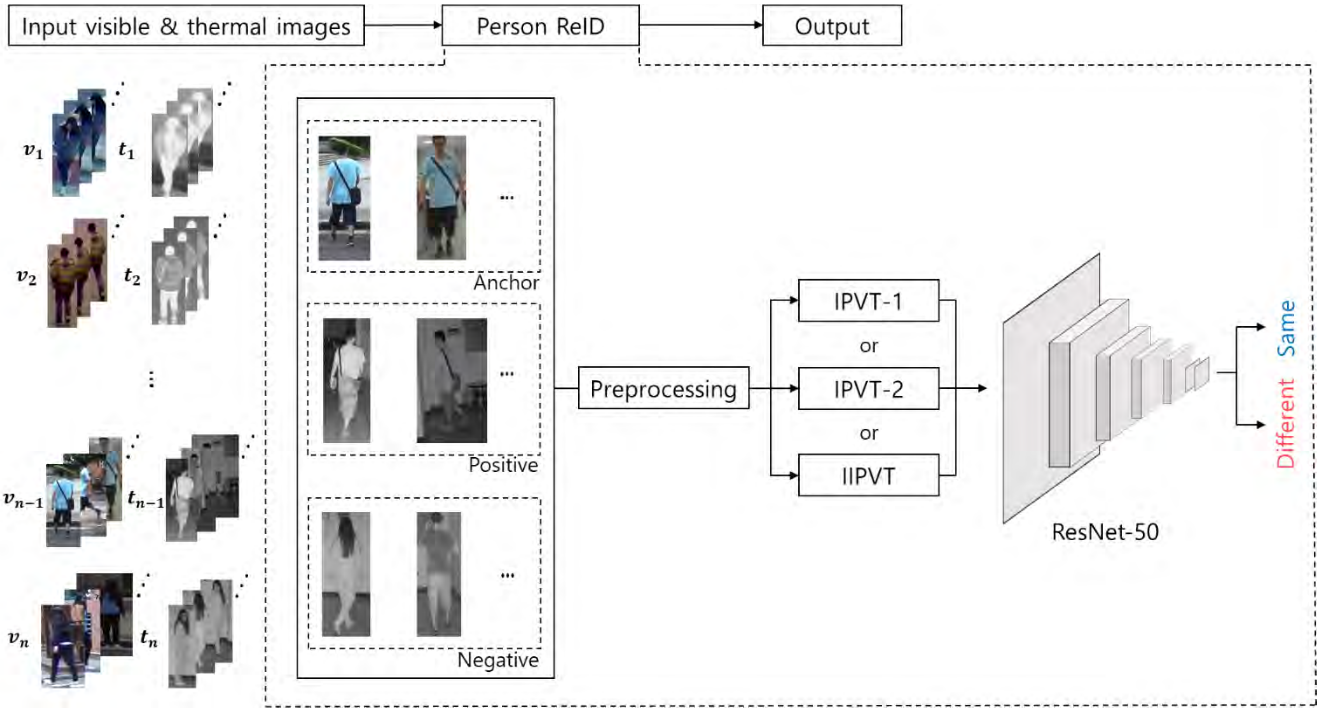


FIGURE 1. Overall procedure of proposed system.

and express the brightness by dividing the image into steps of "0"–"255" for each color. Conversely, the thermal data uses brightness values for the temperature differences in an image by detecting the energy radiated from an object. That is, a visible-light image consists of color and brightness information whereas a thermal image is expressed only with brightness values. To solve the difference of the characteristics of these data, the visible-light image is converted into an image expressed with brightness values using a grayscale only.

Using FaceNet [15] as a reference, the given dataset was divided into anchor, positive, and negative sets. The anchor set was composed of the visible-light image, and the positive and negative sets were composed of thermal images. For each anchor set image, the positive and negative pairs were composed by pairing with the positive and negative set images. Here, inter-channel and intra-channel pairs were composed using the visible-light and thermal images in the positive pair and negative pair. An inter-channel pair is a pair of images that is composed as a single image by arranging the visible-light and thermal images in different channels. Unlike the inter-channel pair, an intra-channel pair refers to images where the visible-light and thermal images in a single channel have a "1 × 2" matrix shape. Using the inter-channel and intra-channel pairs that have undergone this process, the IPVT-1, IPVT-2, and IIPVT were created. As indicated in Fig. 2, IPVT-1 uses the inter-channel pair only; IPVT-2 is an image that is composed of the intra-channel pair only. IIPVT refers to an image that is composed using both the inter-channel and intra-channel pairs. Because the IPVT-1, IPVT-2, and IIPVT contain the data of both the visible-light and thermal images, the learning of personal features appearing commonly in two images is possible when training the CNN; then, from this, the similarity and dissimilarity factors of the two images can be calculated. Here, a method of optimizing the data similarity between the visible-light and thermal images is applied through a preprocessing such as image thresholding or Retinex filtering. Furthermore, because two types of data are expressed in a single image, the learning can be performed using one stream CNN.

1) CASE-I: USING BINARIZED IMAGE

By adding a binary image in the IPVT-1, IPVT-2, and IIPVT models introduced in Section IV.*B*, a more robust person ReID method is proposed. In general, a thermal image tends to express the foreground area bright when the background area is dark; or the foreground area dark when the background area is bright. Therefore, a method to improve the robustness of the matching model proposed in this paper is to classify the foreground and background by applying the Otsu thresholding method [36] to the thermal image. Otsu thresholding is a method that obtains a threshold by identifying a value that maximizes the variance of inter-class under the assumption that the image has a bi-modal histogram [36]. It begins with an assumption that if a binarized image, which is obtained by applying the Otsu thresholding to the thermal image, is added in the input structure of Fig. 2, the learning can be performed in such a manner that the outline of the target object can be

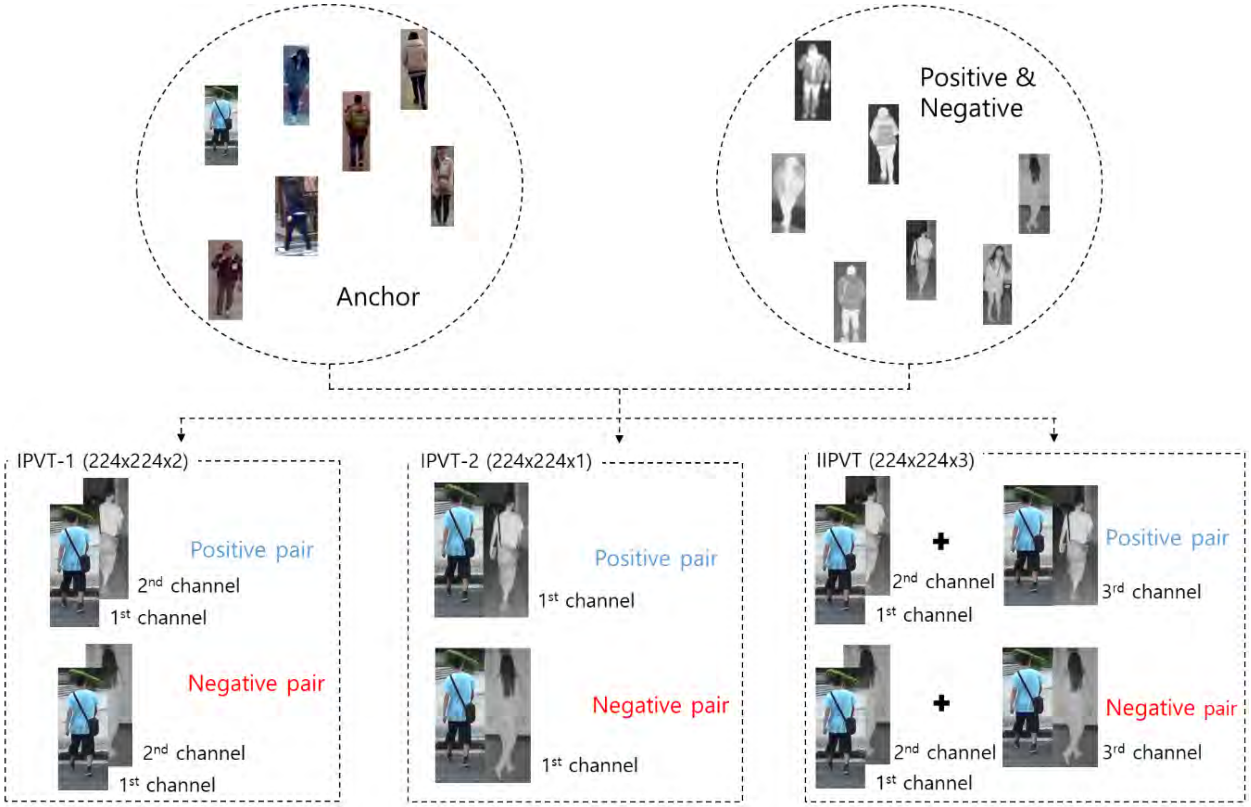


FIGURE 2. Structure of IPVT-1, IPVT-2, and IIPVT.

addressed more sensitively in the deep-feature learning stage. This is because when sensitive to the outline of a target object, the positive and negative pairs can be separated more easily in the case of a database that has no change in the viewpoint.

2) CASE-II. USING RETINEX FILTERED IMAGE

As described in Section IV.*B*, the IPVT-1, IPVT-2, and IIPVT models, which match an RGB image with thermal data using the grayscale brightness information only, was introduced. However, in this section, a method is proposed to maximize the use of the features of the RGB data using the Retinex filter. Retinex theory states that the information of the light source influences the visible information of the object when recognizing the visible information. Land et al. proposed a Retinex algorithm for a visual model that reduces the effect of illumination and compensates to display the object's own color [35]. Retinex is classified into single-scale Retinex (SSR) and multi-scale Retinex (MSR); the SSR is displayed by the following Equation (1).

$$R_i(x, y) = \log I_i(x, y) - \log [F(x, y) * I_i(x, y)] \qquad (1)$$

In Equation (1), $R_i(x, y)$ denotes "true color" or reflectance for each pixel of the input image, which is a result of SSR; $I_i(x, y)$ denotes the input image and $F(x, y)$ denotes the Gaussian filter. That is, it aims to obtain $R_i(x, y)$ from the input image $I_i(x, y)$. By multiplying the weight w_n to $R_i(x, y)$, the MSR is obtained, as indicated in Equation (2) [43]. N is the number of processed image by SSR.

$$R_{MSR_i} = \sum_{n=1}^{N} w_n R_{n_i}(x, y) \qquad (2)$$

C. HETEROGENEOUS FEATURE LEARNING BY CNN

1) ReID BASED ON CNN OUTPUT

In this study, CNN is used to determine whether persons captured by visible-light and thermal cameras are the same person. As indicated in Fig. 1, the learning is performed using the IPVT-1, IPVT-2, or IIPVT composed for every person captured from the visible-light and thermal cameras, as an input for the CNN. For the CNN structure, the ResNet-50 [16] model is used. Moreover, according to the goal of this study, the number of nodes of the fully connected (FC) layer was adjusted, and the pre-trained model was fine-tuned using the experimental data used in this study. The CNN structure used in this study is displayed in Fig. 3 and Table 2. As indicated in Table 2, the number of output nodes of the FC layer in the ResNet-50 structure was modified from 1,000 to two. This was because, as described in Section IV. *B*, the IPVT-1, IPVT-2, and IIPVT were all composed of two classes, i.e., positive and negative pairs. By performing the binary classification in this manner, the similarity and dissimilarity factors for the input image (IPVT-1, IPVT-2, or IIPVT) are produced as the final softmax layer's output.

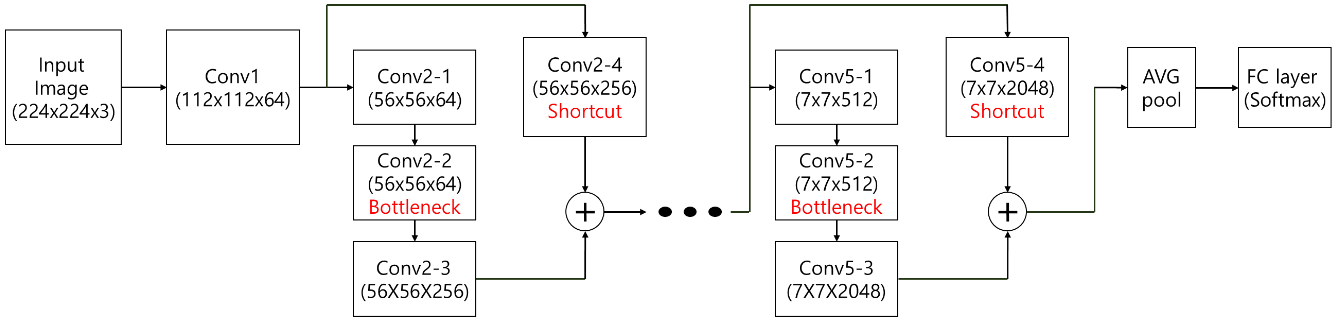


FIGURE 3. Structure of ResNet50.

TABLE 2. ResNet50 architecture (2/1∗ means that the number of strides is two in the first iteration, and one thereafter).

Layer type		Number of filters	Size of feature map (width × height × channel)	Size of filter	Stride	Padding	Number of Iterations
Image input layer			224 × 224 × 3				
Conv1		64	112 × 112 × 64	7×7×3	2	3	1
Max pooling layer		1	56 × 56 × 64	3×3	2	0	1
Conv2	Conv2-1	64	56 × 56 × 64	1 × 1 × 64	1	0	3
	Conv2-2 (Bottleneck)	64	56 × 56 × 64	3 × 3 × 64	1	1	
	Conv2-3	256	56 × 56 × 256	1 × 1 × 64	1	0	
	Conv2-4 (Shortcut)	256	56 × 56 × 256	1 × 1 × 64	1	0	
Conv3	Conv3-1	128	28 × 28 × 128	1 × 1 × 256	2/1*	0	4
	Conv3-2 (Bottleneck)	128	28 × 28 × 128	3 × 3 × 128	1	1	
	Conv3-3	512	28 × 28 × 512	1 × 1 × 128	1	0	
	Conv3-4 (Shortcut)	512	28 × 28 × 512	1 × 1 × 256	2	0	
Conv4	Conv4-1	256	14 × 14 × 256	1 × 1 × 512	2/1*	0	6
	Conv4-2 (Bottleneck)	256	14 × 14 × 256	3 × 3 × 256	1	1	
	Conv4-3	1024	14 × 14 × 1024	1 × 1 × 256	1	0	
	Conv4-4 (Shortcut)	1024	14 × 14 × 1024	1 × 1 × 512	2	0	
Conv5	Conv5-1	512	7 × 7 × 512	1 × 1 × 1024	2/1*	0	3
	Conv5-2 (Bottleneck)	512	7 × 7 × 512	3 × 3 × 512	1	1	
	Conv5-3	2048	7 × 7 × 2048	1 × 1 × 512	1	0	
	Conv5-4 (Shortcut)	2048	7 × 7 × 2048	1 × 1 × 1024	2	0	
AVG pool		1	1 × 1 × 2048	7 × 7	1	0	1
FC layer			2				1
Softmax			2				1

A brief description of the CNN structure is as follows. In the Conv1 layer, a feature map of 112 × 112 × 64 size is generated through convolution computation using 64 filters of size 7 × 7 × 3 in an input image of size 224 × 224 × 3. The size of the feature map generated from the convolution computation can be calculated using the equation, size of feature map = (size of input – size of filter + (2 × padding))/stride + 1 [17]. Conv2–Conv5 are four convolutional layers that contain shortcuts with a bottleneck structure [16].

For the activation function, the rectified linear unit (ReLU) layer was used. The ReLU is displayed in Equation (3) [18]–[20].

$$y = \max(0,x) \tag{3}$$

where x and y are input and output values, respectively. ReLU supplements the vanishing gradient problem [21] that sigmoid produces in the back-propagation for training, and has a faster processing speed than a nonlinear activation function. For this reason, the ReLU has excellent training efficiency compared to other activation functions. In the FC layer, a softmax function [19] similar to Equation (4) is used to calculate the similarity factor of the input image.

$$\sigma(s)_j = \frac{e^{sj}}{\sum_{n=1}^{K} e^{sn}} \tag{4}$$

Given that the array of the output neurons is set as s, the probability of neurons belonging to the jth class is obtained by dividing the value of the jth element by the summation of the values of all the elements.

V. EXPERIMENTAL RESULTS

A. EXPERIMENTAL DATA AND TRAINING

To comparatively evaluate the method proposed in this paper with the person ReID methods of previous studies, two open databases, the DBPerson-Recog-DB1 [22] and SYSU-MM01 [23] databases were used. DBPerson-Recog-DB1 is a database of images that captures front, back, and both sides of the same person with a visible-light (Logitech C600 [40]) and thermal camera (FLIR Tau2 [41]) for 412 people (254 males and 158 females). For each person, there are ten visible-light images of 100 × 110 × 3 pixel size on average and ten thermal images of 110 × 125 × 1 pixel size on average, totaling 8,240 images. SYSU-MM01 is a database of images of different views of each person for 491 people in different brightness, indoors and outdoors, using two infrared light (IR) (thermal) cameras and four visible-light cameras.

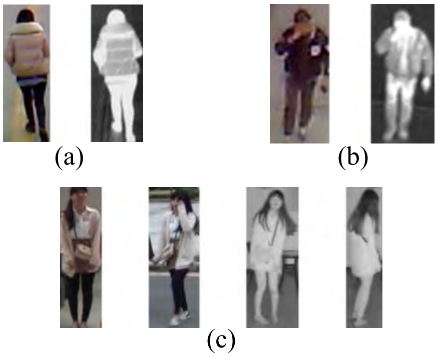

(a)
(b)
(c)

FIGURE 4. Sample images of DBPerson-Recog-DB1 and SYSU-MM01. (a), (b) Sample images of DBperson-Recog-DB1 and (c) sample images of SYSU-MM01. In (a) and (b), left image is a visible-light camera image and right image is a thermal camera image. In (c), two left-side images are visible-light camera images and two right-side images are IR (thermal) camera images.

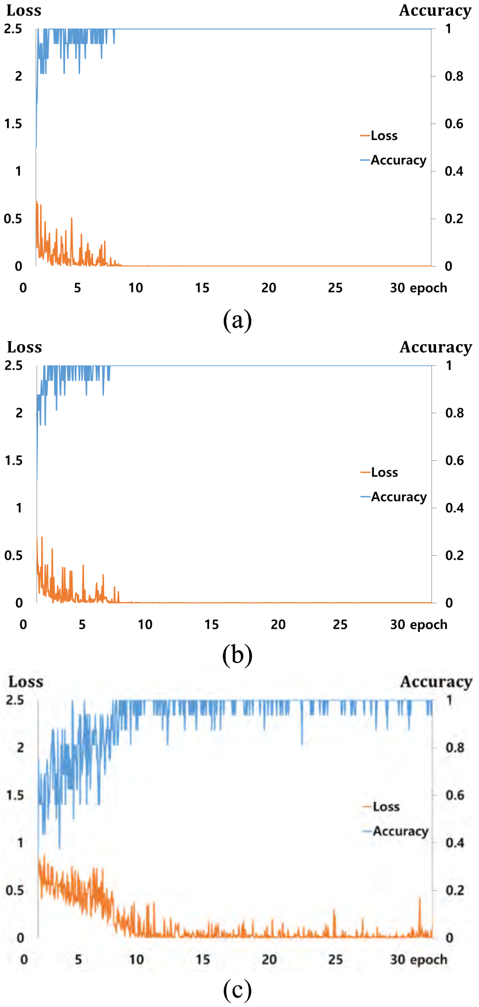
Fig. 4 displays sample images of the DBperson-Recog-DB1 and SYSU-MM01 databases used to verify the proposed method. Figs. 4(a) and (b) are sample images of DBPerson-Recog-DB1. DBPerson-Recog-DB1 is a database that was acquired in an environment set in such a manner that the visible-light and thermal cameras would have the same viewpoint as a single camera. Therefore, as indicated in Figs. 4(a) and (b), for a person, the visible-light and thermal images were captured in the same pose and view. Fig. 4(c) displays sample images of SYSU-MM01, a database that has more diverse parameters than the DBperson-Recog-DB1. As indicated in Fig. 4(c), images were captured in different poses and views for one person. Therefore, SYSU-MM01 is a database that contains the maximum number of parameters for ReID based on visible-light and thermal images.

TABLE 3. Description of database.

	DBPerson-Recog-DB1		SYSU-MM01		
	Subset1	Subset2	Training set	Validation set	Test set
# of images	4,120	4,120	30,213	3,954	10,578
# of class	206	206	296	99	96

To verify the proposed person ReID method, DBPerson-Recog-DB1 was divided into two subsets as indicated in Table 3, and the training and testing were performed in a two-fold cross validation method; for SYSU-MM01, the training set, validation set, and test set were used as they were, as proposed in [23] for a fair comparison. For the CNN training, 41,200 images of the IPVT-1, IPVT-2, and IIPVT per subset of the DBperson-Recog-DB1were used as input images; 187,292 images were used in the SYSU-MM01. The experimental environment used in this study was as follows. All experiments were performed on a desktop computer composed of an Intel® Core$^{\text{TM}}$i7-3770K CPU @ 3.50 GHz (4 CPUs), main memory of 16 GB, and an NVIDIA GeForce GTX 1070 (1,920 CUDA cores) graphics card with memory of 8 GB [34]; the algorithms for the CNN training and testing were implemented by Window Caffe (version 1) [24].

In this study, the stochastic gradient descent (SGD) method [25], [42] was selected as the optimization method for the CNN training. SGD is a method of determining an optimal weight that minimizes the difference of calculated output and desired output based on the computing derivative. Unlike the gradient descent (GD) method, SGD sets one iteration that performs the CNN training with data having a mini-batch size in the training set, and one epoch that repeats the iteration a number of times equal to the training-set size divided by the mini-batch size. The CNN training parameters of this study were as follows: base learning rate of 0.01, learning rate policy of multistep, step value of 5 epochs, maximum iteration of 30 epochs, momentum of 0.9, gamma of 0.1, and weight decay of 0.001. The detail explanations of these parameters can be found in [24]. When the CNN training was performed in the above experimental environment, it consumed eight hours with DBPerson-Recog-DB1 and 35 hours with SYSU-MM01.



(a)

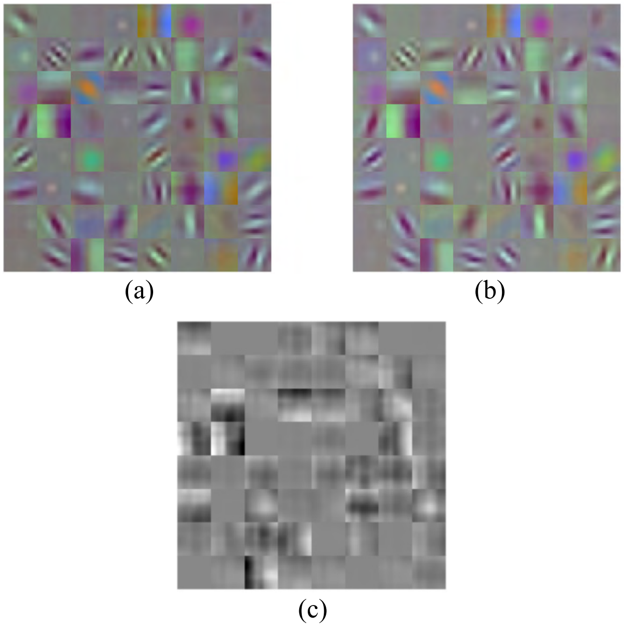
(b)

(c)

FIGURE 5. Loss and accuracy graphs of training procedure: (a) first fold with DBPerson-Recog-DB1, (b) second fold with DBPerson-Recog-DB1, and (c) SYSU-MM01.

Fig. 5 displays the loss and training accuracy graphs in the IIPVT CNN training process based on the number of iterations. In the graphs, the DBPerson-Recog-DB1 and SYSU-MM01 results indicate that as the number of iterations increases, the loss converges to zero and the training accuracy converges to one (100%). Hence, from Fig. 5 it can be confirmed that the CNN has been sufficiently trained.

Fig. 6 displays the 64 filters obtained from the first convolution layer (Conv1 of Table 2) using the two databases. These are the filters extracted when IPVT1-1 was used as the input for DBPerson-Recog-DB1 and IPVT-2 for SYSU-MM01 in the model that demonstrated the highest performance in each database among the three image-pair structures proposed in Section IV.*B*. In Figs. 6(a) and (b), there is virtually no difference in the filter shapes; differences are indicated in the brightness. Because the filter shapes are virtually the same despite the fact that learning was performed for different IDs by dividing the DBPerson-Recog-DB1 into two folds, it can be concluded that the IPVT-1 for ReID is an appropriate model for the DBPerson-Recog-DB. In the case of Fig. 6(c), the majority of the filters indicate an upper/lower and left/right symmetry shape, unlike the cases of (a) and (b). Therefore, it can be stated that the characteristics of the IPVT-2 structure are reflected properly. The reason why the image of Fig. 6(c) is shown as gray is that the input to IPVT-2 is composed of two gray images from visible light and thermal cameras.



(a)

(b)

(c)

FIGURE 6. Example of 64 filters obtained from the first convolution layer (Conv1 of Table 2): (a) first fold, (b) second fold with DBPerson-Recog-DB1, and (c) SYSU-MM01.

B. TESTING OF PROPOSED METHOD WITH DBPERSON-RECOG-DB1

Using the DBPerson-Recog-DB1, the comparative test for the three proposed methods, test for selecting an appropriate model between shallow and deep networks, and comparative test with the methods of the previous studies were performed. for the first experiment, the IPVT-1, IPVT-2, and iipvt proposed in this paper were comparatively tested. for the experimental performance evaluation, the dbperson-recog-db1 was divided into two subsets, as indicated in table 3, and each subset was divided into anchor, positive, and negative sets. moreover, for each anchor image, a positive and negative pair was composed by pairing a positive and negative image. from this, IPVT-1, IPVT-2, and iipvt were generated for the dbperson-recog-db1. the experimental evaluations were compared using rank 1, rank 10, rank 20, and mean average precision (mAP). rank n is a concept that calculates the correct matching accuracy for a case including data (true positive data) of the same class among n matched candidates. map is a mean of the average precision for each query [37]. average precision indicates an area of the precision-recall graph and is an index that evaluates the recognition algorithm. map can be calculated by Equation (5).

$$\mathrm{mAP} = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{5}$$

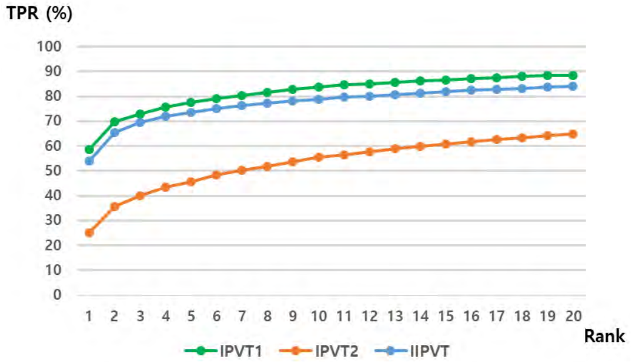


FIGURE 7. Comparative accuracies of IPVT1, IPVT2, and IIPVT with DBPerson-Recog-DB1.

TABLE 4. Comparison of IPVT-1, IPVT-2, and IIPVT with DBPerson-Recog-DB1 (unit: %).

	Rank 1	Rank 10	Rank 20	mAP
IPVT-1	58.57	83.74	88.47	49.11
IPVT-2	25.15	55.46	64.78	25.33
IIPVT	53.96	78.79	84.20	45.01

In Equation (5), Q is the number of queries and $AveP(q)$ is the average precision scores for each query. the experimental results are displayed in Fig. 7 and Table 4. As indicated in Fig. 7, the IPVT-1 always demonstrated excellent performance on the DBPerson-Recog-DB1. in actual measurement values, the IPVT-1'S rank 1 was 58.57%, rank 10 was 83.74%, and rank 20 was 88.47%. As indicated in Table 4, THE mAP OF IPVT-1 was also the highest with a value of 49.11%.

The second experiment was performed to determine the more suitable network between a shallow and deep network using the IPVT-1 as representative of our methods. The shallow and deep networks used in the experiment were AlexNet [3] and ResNet-50, which have proven their performance in the ILSVRC. The corresponding experiment was verified with a two-fold cross validation method using DBPerson-Recog-DB1. The experimental results are presented in Table 5. In Table 5, rank1, rank10, rank20, and mAP all indicate that the deep network provides superior performance compared the shallow network.

TABLE 5. Comparison with AlexNet and ResNet-50 by IPVT-1 on DBPerson-Recog-DB1 (unit: %).

	Rank 1	Rank 10	Rank 20	mAP
Shallow network	50.90	78.40	84.30	41.62
Deep network	58.57	83.74	88.47	49.11

TABLE 6. Comparison with softmax and Euclidean distance on DBPerson-Recog-DB1 (unit: %).

	Rank 1	Rank 10	Rank 20
Softmax	58.57	83.74	88.47
Euclidean distance	0.70	10.07	19.37

TABLE 7. Comparison with Original, Otsu-thresholding, and MSR using IPVT-1 on DBPerson-Recog-DB1 (unit: %).

	Rank 1	Rank 10	Rank 20	mAP
Case 1	58.57	83.74	88.47	49.11
Case 2	50.78	75.49	81.67	41.98
Case 3	58.76	85.75	90.27	47.85

The third experiment compared the proposed method of extracting the similarity of an image pair with softmax in a trained network and the Euclidean distance-based matching using the features extracted from FC layer. The experimental results demonstrate that the softmax method proposed in this paper provided excellent performance overall, as displayed in Table 6.

The fourth experiment aimed to identify preprocessing that could minimize the difference of the image characteristics between the visible-light and thermal images when composing an image pair. The comparative test was performed with a pair composed of grayscale-transformed original visible-light RGB + thermal images (Case 1), a pair composed of binarized images of original RGB, where outline information was added through Otsu-thresholding + thermal images (Case 2), and a pair composed of converted images of original RGB by MSR + thermal image (Case 3). The experiment was performed with reference to the IPVT-1 and deep network (ResNet-50), which demonstrated the best performance in the earlier experiments. The experimental results are displayed in Table 7. When the MSR was applied, the best performance was indicated at rank1 (58.76%), rank 10 (85.75%), and rank 20 (90.27%), and in the case of the original grayscale, mAP (49.11%) was the highest.

Figs.8 and 9 are sample images of the experimental results for IPVT-1 with the best performance on DBPerson-Recog-DB1. Fig. 8 displays a case of obtaining an image for the same person with rank 1, and Fig. 9 displays samples for a person's image calculated with rank 3 and rank 16, respectively. as indicated in Fig. 8, despite the fact that the visible-light and thermal images have large differences in characteristics and image texture, the method of this study demonstrated the correct recognition results.

In Fig. 9, however, the view, clothes, and poses of the images are very similar in spite of negative pairs, which causes the incorrect recognition. actually, the similarity

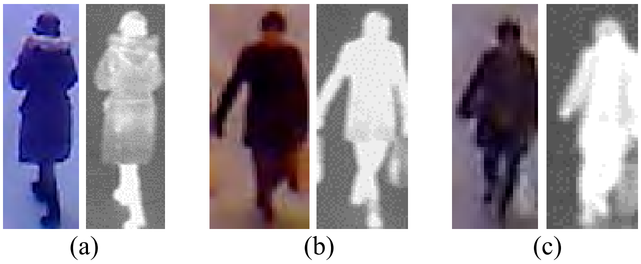


FIGURE 8. Examples of correct ReID test result on DBPerson-Recog-DB1 with: (a) Case 1, (b) Case 2, and (c) Case 3. (a)–(c) Left side image is probe image and right side is rank 1 of gallery images.

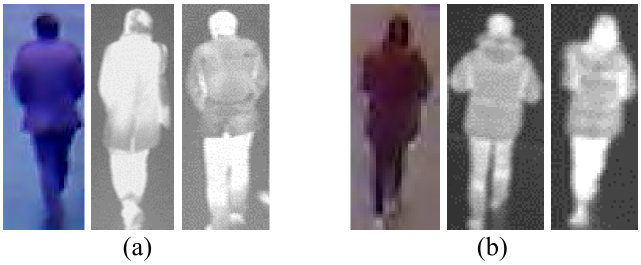


FIGURE 9. Examples of incorrect ReID test result on DBPerson-Recog-DB1: (a) example recorded with rank 3 (left image is probe image, middle image is negative image of gallery set recorded with rank 1, and right image is positive image of gallery set recorded with rank 3) and (b) example recorded with rank 16 (left image is probe image, middle image is negative image of gallery set recorded with rank 1, and right image is positive image of gallery set recorded with rank 16).

factors of the positive and negative pairs were similar, but that of the latter case was a little higher than the former case.

The fifth experiment comparatively evaluated the method proposed in this paper with the methods proposed in the conventional studies (histogram of oriented gradient (HOG) [28], multi-scale local binary patterns (MLBP) [22], local maximal occurrence (LOMO) [26], generalized similarity measure (GSM) [27], zero-padding [23], two-stream cnn network feature learning-hierarchical cross modality metric learning (TONE + HCML) [14], and bi-directional dual-constrained top-ranking (BDTR) [13]). the experimental results are provided in Table 8. As can be observed in Table 8, the IPVT-1 and msr methods that provided the best performances for DBPerson-Recog-DB1 among our methods demonstrated excellent performance compared to the previous methods with respect to rank 1, rank 10, rank 20, and mAP.

TABLE 8. Comparisons with proposed method and previous research with DBPerson-Recog-DB1 (unit: %).

Method	DBperson-Recog-DB1			
	Rank 1	Rank 10	Rank 20	mAP
HOG [28]	13.49	33.22	43.66	10.31
MLBP [22]	2.02	7.33	10.90	6.77
LOMO [26]	0.85	2.47	4.10	2.28
GSM [27]	17.28	34.47	45.26	15.06
Zero-padding [23]	17.75	34.21	44.35	18.90
TONE + HCML [14]	24.44	47.53	56.78	20.80
BDTR [13]	33.47	58.42	67.52	31.83
Proposed (IPVT-1 and MSR)	58.76	85.75	90.27	47.85

TABLE 9. Comparison of IPVT-1, IPVT-2, and IIPVT with SYSU-MM01 database (unit: %).

	Rank 1	Rank 10	Rank 20	mAP
IPVT-1	21.29	48.52	63.34	21.25
IPVT-2	19.95	49.60	65.23	21.72
IIPVT	20.22	43.67	57.68	19.30

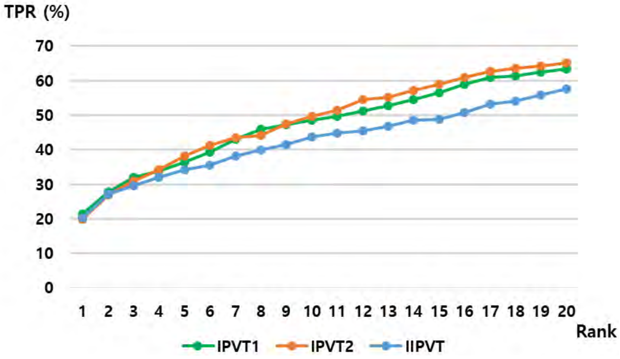


FIGURE 10. Comparative accuracies of IPVT1, IPVT2, and IIPVT with SYSU-MM01 database.

C. TESTING OF PROPOSED METHOD WITH SYSU-MM01 DATABASE

To verify the method proposed in this paper with an open data with varying parameters, the SYSU-MM01 database was used. An experiment comparing the IPVT-1, IPVT-2, and IIPVT and a comparative experiment with the methods of person ReID proposed by previous researchers was performed. The experiments were conducted by dividing the SYSU-MM01 database into a training set, validation set, and test set, similar to the conditions used in the experiment of a previous researcher [23]. Using the images classified into three sets, the IPVT-1, IPVT-2, and IIPVT were generated. The experimental evaluations used rank 1, rank 10, rank 20, and mAP, similar to those of DBPerson-Recog-DB1. The results of the first experiment are displayed in Fig. 10 and Table 9. In Fig. 10, it can be confirmed that the matching rate of IPVT-2 was superior overall compared to the IPVT-1 and IIPVT. In the actual measurement values, the IPVT-1 was 21.29% at rank 1, the highest value at this rank, and IPVT-2 was 49.60% and 65.23% at the rank 10 and rank 20, respectively, the highest values at these ranks. In Table 9, mAP is 21.72% for IPVT-2, indicating higher performance than the IPVT-1 and IIPVT.

The second experiment performed a comparison with MSR, a preprocessing method verified on DBPerson-Recog-DB1 previously. The experiment compared the accuracy of

TABLE 10. Comparison with original and MSR by IPVT-2 on SYSU-MM01 (unit: %).

	Rank 1	Rank 10	Rank 20	mAP
Case 1	19.95	49.60	65.23	21.72
Case 3	23.18	51.21	61.73	22.49

the methods of Cases 1 and 3 that demonstrated excellent performance among Cases 1–3 described in Table 7. The image pairs were based on the IPVT-2 optimized for the SYSU-MM01 database as verified in Table 9. The experimental results are displayed in Table 10. In Table 10, the rank 1, rank 10, and mAP, excepting rank 20, demonstrate that the MSR method (Case 3) provides excellent performance. Through these experimental results, it can be confirmed that the MSR (Case 3) is more favorable for matching (rank concept accuracy) with the thermal data than the gray-scaled RGB data (Case 1), as was the case in the experiment on DBPerson-Recog-DB1.

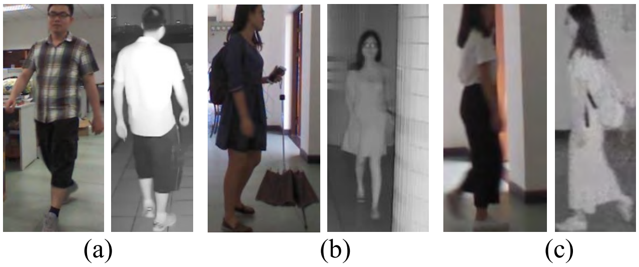


FIGURE 11. Examples of correct ReID test result on SYSU-MM01 database with: (a) Case 1, (b) Case 2, and (c) Case 3. (a)–(c) Left image is probe image and right image is rank 1 of gallery images.

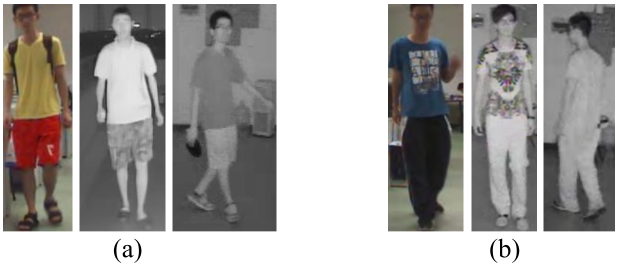


FIGURE 12. Examples of incorrect ReID test result on SYSU-MM01 database: (a) example recorded with rank 3 (left image is probe image, middle image is negative image of gallery set recorded with rank 1, and right image is positive image of gallery set recorded with rank 3) and (b) example recorded with rank 16 (left image is probe image, middle image is negative image of gallery set recorded with rank 1, and right image is a positive image of gallery set recorded with rank 16).

Figs. 11 and 12 are sample images of the experimental results for IPVT-2 that demonstrate the best performance on the SYSU-MM01 database. Fig. 11 exhibits a case of calculating an image for the same person with rank 1, and Fig. 12 displays examples of calculating a person's image with rank 3 and rank 16, respectively. As indicated in Fig. 11, although there were large differences in the image's characteristics, target object size, view, and texture between the visible-light and thermal images, the method of this study produced correctly recognized results.

In Fig.12, the clothes and poses of probe are similar to those in both negative and positive pairs. However, the view of probe image is much similar to that of negative pair compared to positive pair, which causes incorrect recognition. The third experiment compared the proposed and previous methods. The results are presented in Table 11. In Table 11, it can be observed that among the previous methods, BDTR provided the best performance with 17.01%, 55.43%, 71.96%, and 19.66% for rank 1, rank 10, rank 20, and mAP, respectively. For rank 1 and mAP, the proposed method improved the performance by 6.17% and 2.83%, respectively, compared to the BDTR; however, for ranks 10 and 20, it decreased the performance by 4.22% and 10.23%, respectively. A future project will be to conduct a study on a performance improvement method for rank 10 and rank 20.

TABLE 11. Comparisons with proposed method and previous research with DBPerson-Recog-DB1 (unit: %).

Method	SYSU-MM01			
	Rank 1	Rank 10	Rank 20	mAP
HOG [28]	2.76	18.25	31.91	4.24
MLBP [22]	2.12	16.23	28.32	3.86
LOMO [26]	1.75	14.14	26.63	3.48
GSM [27]	5.29	33.71	52.95	8.00
Zero-padding [23]	14.80	54.12	71.33	15.95
TONE + HCML [14]	14.32	53.16	69.17	16.16
BDTR [13]	17.01	55.43	71.96	19.66
Proposed (IPVT-1 and MSR)	23.18	51.21	61.73	22.49

D. DISCUSSIONS

When Tables 8 and 11 are compared, it can be confirmed that the performance indicates a difference between the DBPerson-Recog-DB1 and SYSU-MM01 databases. The reason for this is as follows. For DBPerson-Recog-DB1, because the visible-light and thermal cameras captured the images of the target object at similar locations, the view difference in the two camera images were not large, as illustrated in Figs. 4(a) and (b). Conversely, in the SYSU-MM01 database, although the target objects photographed with the visible-light and thermal cameras are identical, the viewpoint and surrounding environment are different, as displayed in Fig. 4(c). Therefore, because the SYSU-MM01 database has conditions that are more difficult for the person ReID compared to the DBPerson-Recog-DB1, the difference is reflected in the performance. Furthermore, the favorable image pairs were confirmed according to the camera-shooting environment through the experiment. When the visible-light and thermal cameras were capturing images at a same location, the IPVT-1 was suitable; with different locations or views, the IPVT-2 was suitable.

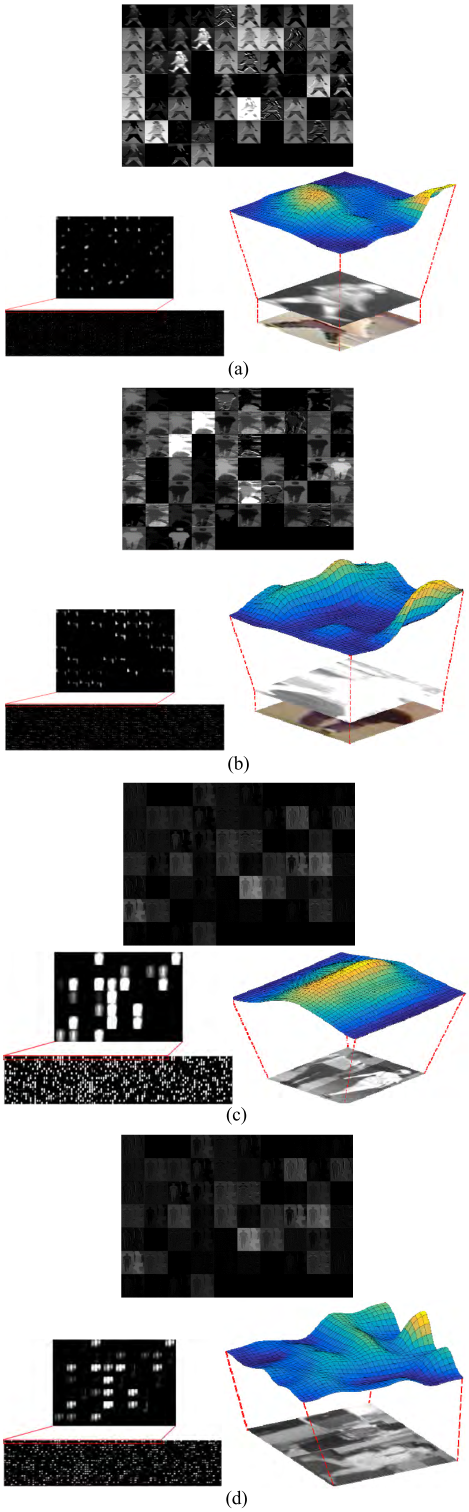


FIGURE 13. Examples of feature maps through tests obtained with (a), (b) DBPerson-Recog-DB1 and (c), (d) SYSU-MM01 database. (a) and (c) are from positive pair whereas (b) and (d) are from negative pairs. In (a)–(d), upper images display the feature maps obtained from first convolution layer (Conv1 of Table 2), whereas lower left images display feature maps obtained from last convolution layer (before AVG pool of Table 2). In (a)–(d), lower right images display 3D feature map image based on average feature map values of lower left images, respectively.

Fig. 13 visualizes the feature maps obtained through the deep CNN used in this study for a single image of the test data (positive pair in case of Figs. 13(a) and (c), negative pair in case of Figs. 13(b) and (d)). As indicated in the lower right images of Figures 13(a)–(d), the change of value in the feature map is relatively large in the case of the negative pair compared to that of the positive pair. From this, it can be observed that the positive and negative pairs can be classified. That is, from Fig. 13, it is confirmed that the image pair-based deep CNN method proposed in this study can be used effectively for person ReID.

VI. CONCLUSION

This paper proposed a method that facilitates person ReID with a one-stream network by composing an inter-channel an intra-channel pair for person regions in images captured from visible-light and thermal cameras. To determine the best possible combinations from the inter-channel and intra-channel pairs, the IPVT-1, IPVT-2, and IIPVT were comparatively evaluated and an image pair that demonstrated the best performance was identified. Regarding DBPerson-Recog-DB1, the IPVT-1 displayed the best performance based on rank 1, rank 10, rank 20, and mAP. Furthermore, it was proven experimentally that the combination with a deep network was more efficient than that with a shallow network when IPVT-1 was used representatively among the three proposed methods. Further, when compared based on the three preprocessing methods, the IPVT-1 and MSR combination demonstrated the best performance. Regarding the SYSU-MM01 database, the IPVT-1 displayed the best performance based on rank 1, and the IPVT-2 provided the performance based on ranks 10 and 20, and mAP. Furthermore, as in the case of DBPerson-Recog-DB1, the MSR preprocessing method proved excellent performance experimentally. Lastly, for a fair comparison with the methods of other researchers, a comparative evaluation was performed with other conventional studies. In the experimental results, the method of combining the MSR with one stream network based on the IPVT-1, IPVT-2, and IIPVT demonstrated superior performance compared to other conventional person ReID methods based on a multimodal camera.

In future, a study will be conducted on a performance improvement method for rank 10 and rank 20 for the case of large variations between respective camera images, similar to those in the SYSU-MM01 database. Furthermore, a study will be performed to reduce the difference between respective camera images through image conversion (creation) using a generative adversarial network (GAN), and through this, improving the performance of person ReID.

REFERENCES

[1] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol, 34, no. 1, pp. 3–19, 2013.

[2] W. Zajdel, Z. Zivkovic, and B. J. A. Krose, "Keeping track of humans: Have I seen this person before?" in *Proc. Int. Conf. Robot. Automat.*, Barcelona, Spain, Apr. 2005, pp. 2081–2086.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.

[4] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 152–159.

[5] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. IEEE 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 34–39.

[6] L. Zheng, Y. Yang, and A. G. Hauptmann. (2016). "Person re-identification: Past, present and future." [Online]. Available: https://arxiv.org/abs/1610.02984

[7] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3908–3916.

[8] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 135–153.

[9] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[10] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun./Jul. 2016, pp. 1335–1344.

[11] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 1320–1329.

[12] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2588–2603, Jun. 2017.

[13] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 1092–1099.

[14] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. 32nd Assoc. Advancement Artif. Intell. Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 7501–7508.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun./Jul. 2016, pp. 770–778.

[17] *CS231n Convolutional Neural Networks for Visual Recognition*. Accessed: Dec. 27, 2018. [Online]. Available: http://cs231n.github.io/convolutional-networks/#overview

[18] *Convolutional Neural Network*. Accessed: Dec. 27, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Convolutional_neural_network

[19] J. Heaton, *Artificial Intelligence for Humans: Deep Learning and Neural Networks*, vol. 3. St. Louis, MO, USA: Heaton Research, 2015.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 807–814.

[21] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA, Apr. 2011, pp. 315–323.

[22] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[23] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 5380–5389.

[24] *Caffe*. Accessed: Dec. 27, 2018. [Online]. Available: http://caffe.berkeleyvision.org

[25] *Stochastic Gradient Descent*. Accessed: Dec. 27, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Stochastic_gradient_descent

[26] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 2197–2206.

[27] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1089–1102, Jun. 2017.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[29] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1386–1393.

[30] A. Møgelmose, C. Bahnsen, T. B. Moeslund, A. Clapes, and S. Escalera, "Tri-modal person re-identification with RGB, depth and thermal features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, Jun. 2013, pp. 301–307.

[31] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[32] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 116–124.

[33] *Dongguk Person ReID CNN Models*. Accessed: Dec. 27, 2018. [Online]. Available: http://dm.dgu.edu/link.html

[34] *Geforce GTX 1070*. Accessed: Dec. 27, 2018. [Online]. Available: https://www.geforce.com/hardware/desktop-gpus/geforce-gtx-1070/specifications

[35] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, pp. 1–11, 1971.

[36] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[37] *mAP*. Accessed: Dec. 27, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)#Mean_average_precision

[38] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[39] T. Huang and S. Russell, "Object identification in a Bayesian context," in *Proc. 15th Int. Joint Conf. Artif. Intell.*, Nagoya, Japan, Aug. 1997, pp. 1276–1282.

[40] *C600 Webcam Camera*. Accessed: Feb. 12, 2018. [Online]. Available: https://support.logitech.com/en_us/product/5869

[41] *Tau2 Thermal Imaging Camera*. Accessed: Feb. 12, 2018. [Online]. Available: http://www.flir.com/cores/display/?id=54717

[42] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. 19th Int. Conf. Comput. Statist.*, Paris, France, Aug. 2010, pp. 177–186.

[43] Z. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-scale retinex for color image enhancement," in *Proc. 3rd IEEE Int. Conf. Image Process.*, Lausanne, Switzerland, Sep. 1996, pp. 1003–1006.

JIN KYU KANG received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2016, where he is currently pursuing the combined course of M.S. and Ph.D. degrees in electronics and electrical engineering. His research interests include biometrics and deep learning.

TOAN MINH HOANG received the B.S. degree in information and communication technology from the Hanoi University of Science and Technology, Hanoi, Vietnam, in 2013. He is currently pursuing the combined course of M.S. and Ph.D. degrees in electronics and electrical engineering with Dongguk University. His research interests include biometrics and deep learning.

KANG RYOUNG PARK received the B.S. and M.S. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1994 and 1996, respectively, and the Ph.D. degree in electrical and computer engineering from Yonsei University, in 2000. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since 2013. His research interests include image processing and biometrics. He supervised this research and helped the revision of the draft of the original paper.