Distance based Training for Cross-Modality Person Re-Identification

Nihat Tekeli, Ahmet Burak Can
Department of Computer Engineering, Hacettepe University
Beytepe, Ankara TR-06800, Turkey
{nihat.tekeli,abc}@hacettepe.edu.tr

Abstract

Cross-modality person re-identification between infrared (IR) and visible (VIS) domains is a challenging problem, which aims to identify persons in different spectrums, variety of camera specs, and broad illumination conditions. This paper proposes distance based training on an one-stream convolutional neural network architecture, in which network weights are shared between IR and VIS domains to learn discriminative features for person re-identification. The distance based score layer enables to train the network using distance metrics instead of the fully connected layer. Different distance metrics can be used for training and ranking stages. The proposed structure enables to extract discriminative features in the cross-modality data without using dedicated structures for each domain. Experimental results on a cross-modality person re-identification dataset indicate that the proposed approach outperforms the state-of-the-art methods.

1. Introduction

Image based person re-identification aims to find occurrences of the same person in an image dataset given a query image. The challenge is to correctly match two images of the same person. Pose variations, human body deformation, occlusion, camera differences, light conditions, background differences make person re-identification a difficult task. In the recent years, many approaches addressing this problem are developed with increasing performance. Early person re-identification works extract hand-crafted features from input images using mostly color information [6, 9, 17, 30]. Person re-identification is done by comparing the feature vectors of a query image and gallery images using a distance metric. More recent works extract feature vectors using convolutional neural network (CNN) architectures, which are mostly pre-trained on another large dataset [2, 4, 12, 23]. These works benefit from discriminative capacity of modern deep learning architectures combined with different loss functions. Commonly used loss functions include identification loss, triplet loss, and other distance constraints on feature vectors.

High majority of the works in the literature focus on daytime person re-identification using VIS cameras [14, 22, 26, 34, 35]. However, matching persons in the night time has the similar importance in surveillance applications. VIS cameras perform poorly and take under exposed images in low light environments. The cameras with near infrared (NIR) capture feature are widely used in dark conditions mostly with NIR illuminators. Cross-modality person re-identification aims to match images of the same individuals in different modalities, *i.e*., IR images taken at dark environments with VIS images taken at well illuminated areas. Since IR images are recorded as 1-channel, they do not contain color information. In the lack of color information, the task becomes more difficult than the person re-identification on standard VIS band cameras. There are only a limited number of studies investigating cross-modality person re-identification. Wu *et al*. [27] introduce SYSU-MM01 dataset and propose a method on NIR-VIS cross-modality person re-identification. Ye *et al*. [31, 32] and Dai *et al*. [5] propose other architectures to improve performance further.

Recent person re-identification approaches generally consists of two stages: i) training a deep neural network, ii) ranking gallery images according to the query image. After the training stage, deep features of the query image and gallery images are obtained by forward passing images through the network. Then, feature vectors of gallery images and the query image are compared using a distance metric. The gallery images with lower distances are returned as results. Generally, the fully connected layer is used in the last layer of the architecture for identification purposes. In addition, contrastive loss and triplet loss methods are widely used to gather samples of the same identity closer.

This paper proposes an approach to use distance metrics in training and ranking stages. An one-stream network architecture with a distance based score layer is proposed to match individuals in infrared (IR) and VIS domains. The fully connected layer of a CNN is replaced with the distance

based score layer in order to use distance metrics in both the training and ranking stages. The architecture relies on the discriminative capability and large number of weights in the network to adapt to different domains. The architecture behaves indifferently between cross-modality inputs. The proposed distance based score layer precedes the loss function and gives higher scores for closer feature vectors and lower scores for distant feature vectors according to the distance metric.

Main contributions of this work can be highlighted as follows:

- An end-to-end one-stream network is shown to have discriminative ability for two different modalities without specific operations on each domain.
- The distance based score layer is proposed to train the network with a distance metric.
- The experimental results are presented on two cross-modality datasets. These results show that the proposed approach outperforms the state-of-the-art methods.

The rest of the paper is organized as follows: Section 2 reviews the related works in person re-identification and cross-modality identification. Section 3 explains the details of the proposed method. Section 4 presents the experimental results. Section 5 outlines the results and findings of the work.

2. Related Work

Image based person re-identification is traditionally carried out using two strategies, which are hand-crafted feature extraction methods and metric learning algorithms. Developing hand-crafted features focus on finding discriminative features that are robust to pose, view, and illumination changes [6, 9, 17, 30]. Metric learning methods focus on learning distance metrics which act on features extracted from person images. Some recent works rely on deep learning methods to extract representative features. In the following paragraphs, a few studies related to deep learning in person re-identification, cross-modality recognition, and cross-modality person re-identification are outlined.

Deep learning methods are widely used in the literature for person re-identification problem. In person re-identification applications, different types of networks are applied such as single convolutional neural networks [3, 23, 29], siamese networks [1, 24, 33, 37], triplet loss network architectures [2, 4, 12], and fisher networks [28]. The general strategy on CNNs for person re-identification is to train a network and extract features by forward passing test images through the trained network. Then, feature vectors of gallery images are ranked by comparing with the feature vector of the query image using a distance metric.

Siamese architectures based on dual or shared convolutional neural networks are used in person re-identification. In [37], a siamese network containing identification and verification loss stages is proposed. In [24], a gating function is proposed to emphasize different local patterns for different image pairs during comparison in a Siamese CNN network.

Many studies apply triplet loss on their network architectures and benefit from hard triplet selection. In [4], the triplet network is used to obtain feature vectors from three input images. The network is trained to reduce distance between similar image pairs and increase the distance between different image pairs. Almazan *et al*. [2] use Residual Networks [10] as the backbone network architecture to implement a triplet loss architecture, which accepts arbitrary image sizes and use random erasing to augment input data. Hermans *et al*. [12] propose a batch based hard triplet selection scheme and show that performance is improved using triplet loss in person re-identification.

Methods that divide image into splits or divide human body into parts are proposed in the literature. Zhang *et al*. [34] divide person images into horizontal stripes and extract aligned features to match person images. Kalayeh *et al*. [14] parse human body semantically and obtain features corresponding to each body part using a deep learning architecture. By matching the body parts, person re-identification performance is improved. Similar to this study, Su *et al*. [22] and Zhao *et al*. [35] extract local features after finding human body parts. In [26], a multiple granularity network architecture is proposed. The architecture extracts local features from horizontal stripes and global features from the whole image.

Cross-modality identification or recognition are also studied in other fields such as iris recognition [19, 25] and face recognition [11, 13, 15, 18]. PolyU NIR-VIS Iris dataset [19] and CASIA NIR-VIS 2.0 Face Recognition dataset [16] are the two cross-modality datasets used in iris and face recognition, respectively. There are a few deep learning studies in cross-modality face identification. He *et al*. [11] propose a two-stream network architecture that finds shared features between different domains using the maxout operator [8] and orthogonal constraints. Liu *et al*. [18] use a triplet loss for face recognition to train a network with convolutional layers, max-feature-maps, and max pooling layers. Despite similarities, cross-modality person re-identification is more challenging than iris/face identification due to large view changes and occlusion. In addition, biometric information is not available in person images taken from surveillance cameras. Body parts, clothing, and belongings are used to identify individuals across different cameras and modalities.

There are a small number of studies that focus on cross-modality person re-identification using deep learning methods. In addition, the datasets available to researchers

are very limited on this problem. SYSU-MM01 NIR-VIS cross-modality person re-identification dataset [27] and RegDB Thermal-VIS cross-modality dataset [20] are the two datasets publicly available. Wu *et al*. [27] propose deep-zero padding network architecture, which uses two different input channels for NIR and grayscale VIS images in order to provide the network with cross-modality person re-identification data. Their method performs better than some known hand-crafted features. Ye *et al*. [31] proposes an architecture consisting of feature learning and metric learning stages. In feature learning stage, two stream network structure is used with identification loss and contrastive loss. In metric learning stage, modality specific and modality shared metrics are studied. In another work of Ye *et al*. [32], dual-constrained top ranking algorithm is used to arrange intra and cross-modality distances for cross-modality person re-identification. The results show improved performance on both Thermal-VIS and NIR-VIS cross-modality datasets. Dai *et al*. [5] use a triplet network with triplet constraints and identification loss to learn discriminative features. In addition, generative adversarial training [7] is used to remove domain specific information in the extracted features. The generative network aims to extract domain invariant features; whereas, the adversarial network aims to find modality of the given feature vector.

3. The Proposed Method

In this section, the proposed deep network architecture based on Residual Networks (ResNet) [10] is explained. In the architecture, the last fully connected layer is replaced with the proposed distance based score layer. In the following subsections, after investigating internals of the fully connected layer, the distance based score layer is introduced. The network structure and training strategy are explained afterwards.

3.1. Investigation of Fully Connected Layer

Most deep learning networks include a fully connected layer at the last stage for classification purposes. The input size of a fully connected layer is determined by the size of the previous layer's output vector. This vector is also called the feature vector in person re-identification applications since it is generally used for image comparison in the ranking stage. The fully connected layer's output size is determined by number of classes in classification applications. A fully connected layer without a bias term can be defined as matrix multiplication, where the first element is a matrix containing weights of the fully connected layer, the second element is the input vector of the fully connected layer. If a bias term is used, an additional term is added to the result of multiplication. In Equation 1, a function defining the operation of a fully connected layer with bias term is given.

$$classScores = f(x, W) = Wx + b \tag{1}$$

where W is the C by N weight matrix. C denotes the number of classes (output vector size) and N denotes the number of input elements (input vector size). x is the input vector with the length of N. The bias term b is a vector of length C. The multiplication operation produces a vector of length C, which gives similarity scores for each class. Equation 2 shows Equation 1 in the expanded form.

$$f(x, W) = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,N} \\ w_{2,1} & w_{2,2} & \dots & w_{2,N} \\ \vdots & \dots & \ddots & \vdots \\ w_{C,1} & w_{C,2} & \dots & w_{C,N} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_C \end{bmatrix} \tag{2}$$

The score value for class i ($Score_{class_i}$) is calculated by multiplying the i^{th} row of the weight vector by the input vector and summing up the multiplication results. A fully connected layer internally defines a center vector for each class *i.e*. there are C center vectors if the output size is C. Score value is generated by using dot product between the input vector and the center point (vector) of each class. Therefore, score is the dot product of the corresponding row of weight vector and the input vector as seen in Equation 3.

$$Score_{class_i} = \begin{bmatrix} w_{i,1} & w_{i,2} & \dots & w_{i,N} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} + b_i \tag{3}$$

Since the class score is determined by calculating dot product, class score is actually the similarity of the input vector and the center vector of the class if bias term is ignored. Another interpretation is that class score is cosine similarity between the input vector and the center vector of the class multiplied by product of their magnitudes.

3.2. The Distance based Score Layer

Inspired by the fully connected layer, the distance based score layer defines center points of classes and outputs score values. While the fully connected layer uses only dot product operations, score calculation is not fixed to a single metric in the distance based score layer, it is possible to use different distance metrics. Figure 1 gives an illustration of how class scores are calculated and how output values are generated in the distance based score layer. As seen in Figure 1, the output of distance function is multiplied by -1. Therefore, class score is the negative of the distance metric calculated in the layer. The reason for using the negative distance value is to make the score value high when the distance between two vectors is small. A high score value for a class means that the given sample is close to the center point of the class. As a result of producing higher scores for

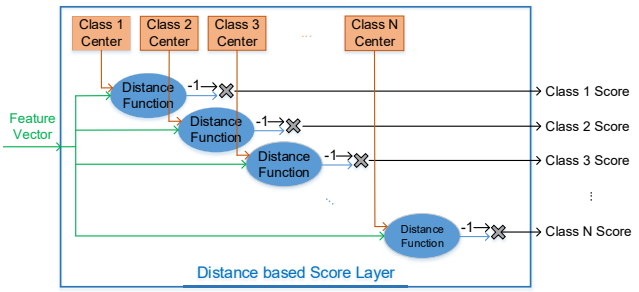

Figure 1. Internal view of the distance based score layer. The distance function calculates the distance between the feature vector and the corresponding class center. Then, the distance is multiplied by -1 to get the class score. Operations are repeated for each output class. For ease of explanation, the distance based score layer is shown with a single feature vector input instead of a batch of feature vectors. However, the distance based score layer can accept a batch of inputs and output a batch of scores.

lower distances in the distance based score layer, it is possible to use standard loss functions in training without any modification.

It is possible to use different distance metrics in the distance based score layer. Manhattan (L1) distance, Euclidean (L2) distance, Euclidean squared distance (L2-Squared), and Chebyshev distance (L∞) metrics are used in this work and compared in the later sections. In addition to distance metrics, negative of dot product between input vectors and center points is used in a similar way to distance metrics. When negative of dot product is used, the distance based score layer performs the same operations as the fully connected layer if no bias term is used and it is expected that the performance of the network should be similar to the fully connected layer. Thus, correct execution of the distance based score layer can be tested. Negative of dot product should give an indication that the distance based score layer works and the network is properly trained if its scores are close to the fully connected layer.

Let x be the input vector and c_i be the class center of class i. Then, L1, L2, L2-squared, L∞ distance functions, and negative of dot product are given in Equations 4, 5, 6, 7, and 8, respectively. These functions are directly used in the distance function block seen in Figure 1. It should be noted that negative of dot product is represented as d_{Dot} and used similar to a distance metric, whereas it is not an actual distance metric.

$$d_{L1}(x, c_i) = \sum_j |x_j - c_{i_j}| \tag{4}$$

$$d_{L2}(x, c_i) = \sqrt{\sum_j (x_j - c_{i_j})^2} \tag{5}$$

$$d_{L2-Square}(x, c_i) = \sum_j (x_j - c_{i_j})^2 \tag{6}$$

$$d_{L\infty}(x, c_i) = \max_j |x_j - c_{i_j}| \tag{7}$$

$$d_{Dot}(x, c_i) = -x \cdot c_i = -\sum_j (x_j c_{i_j}) \tag{8}$$

where x_j and c_{i_j} indicates j^{th} element of x and c_i vectors, respectively.

In the forward pass of the network, the distances between input vectors and center points are calculated and given to output. However, the center points of classes need to be updated after scores are calculated. The standard way to update the network variables is to apply backpropagation starting from the loss value and find gradients. Then, variables are updated with an algorithm which uses gradients, such as stochastic gradient descent. Using a gradient update algorithm, center points are gradually shaped and updated. Therefore, the network is trained to make samples of the same class closer according to a specific distance metric.

The distances are multiplied by -1 to convert them to negative value and it must be taken into account for backpropagation as well. When the partial derivatives of distance functions are examined, it is seen that gradients depend only on x_j and c_{i_j} values for L2-Squared distance metrics and negative of dot product. However, partial derivative of L2 distance depends on other elements of vectors x and c_i due to square-root term. Center points can not be quickly updated when L2 distance is used since update of each element depends on update of other elements of the center point during training. Hence, using backpropagation and gradient update scheme might not lead to convergence in a reasonable amount of time for L2 distance metric. An update scheme different than gradient update methods is needed for L2 distance metric. To find center points for L2 distance and decrease convergence time, the center points of each class is calculated by taking the exponential running average of feature vectors. The update equation of the center points of classes are given in Equation 9.

$$c_{label_x}^{t+1} = \beta c_{label_x}^{t} + (1 - \beta) x^t \tag{9}$$

where β is a hyperparameter that affects center point update rate. $label_x$ denotes the label (target) of input vector x. While superscript t indicates the current value, $t + 1$ indicates the next value. The exponential running average method continuously updates center points as weights of the convolutional neural network get trained with backpropagation and Stochastic Gradient Descent (SGD). In addition, update rate of center points are controlled with parameter β, which should be selected carefully to get stable and fast enough updates similar to selection of learning rate.

In adddition to gradient update and exponential running average, median and mid-range of feature vectors can be

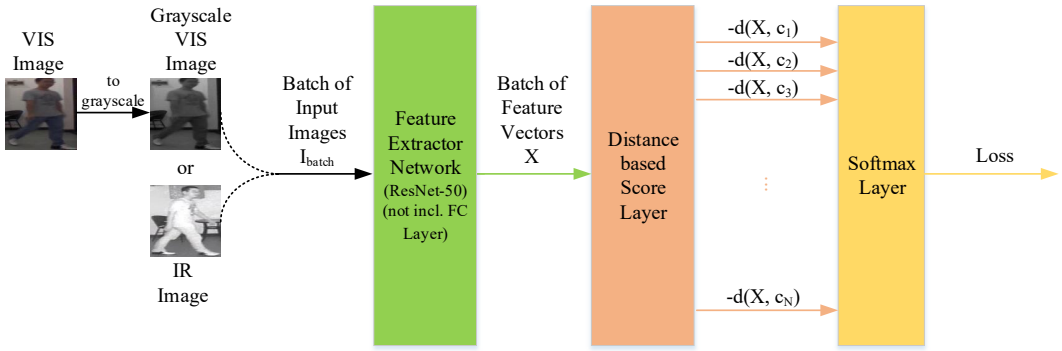

Figure 2. The proposed network architecture with the distance based score layer. I_{batch} is a batch of one-channel images containing IR and VIS data. X is batch of feature vectors and is input to the distance based score layer. $d(X, c_j)$ is batch of distances between batch of feature vectors and center point of j^{th} class.

used to update center points. Median of a set is the middle element in sorted form of the set. Mid-range is the arithmetic mean of maximum and minimum elements of a given set. Basically, median value and mid-range value of the feature vectors in the batch are substituted with center points.

The above mentioned distance metrics and center update schemes are tested in Section 4.

3.3. Network Architecture

The proposed network architecture is constructed upon ResNet-50 base architecture [10]. The network is initialized using the pre-trained ResNet-50 model and fine-tuned on person re-identification task. The reason for using pre-trained network is that ResNet model is trained on 1.4 million images containing large variety of shapes and textures [21] but the cross-modality datasets used in the experiments contain only tens of thousands of images. Large networks require large datasets to adapt to the desired structure and generalize for real world tasks. However, most datasets available for person re-identification is not sufficient to train the ResNet model from scratch. Many studies use the pre-trained ResNet model and thus the same methodology is followed in this work as well.

Modifications are carried out in the first layer and last few layers of the network to adapt it for cross-modality training. The first layer of ResNet model is substituted with an one-channel convolutional layer to accept IR and VIS images. In order to forward IR and VIS images in the same way, VIS images are converted to one-channel grayscale images. IR and VIS images are chosen randomly in a batch to be forward-passed through the network. Furthermore, output size of the last fully connected layer is changed with the number of persons in the training set of the cross-modality dataset. In addition, the last fully connected layer is replaced with the proposed distance based score layer in the related experiments. Model is trained to identify individuals via output of softmax loss function.

During the development stage, input images with 3-channel and 1-channel are tested. In 3-channel setting, the first layer of the network is not changed and RGB images are directly fed into the network. The single channel IR images are duplicated to construct 3-channel images. No significant performance difference is observed in these settings and thus 1-channel setting results are reported in Section 4 due to space limitations.

The output of average pooling layer of ResNet-50 is accepted as the feature vector for a given person image. This feature vector is an one-dimensional vector of size 2048 for ResNet-50 architecture. Irrespective of input image modality, the feature vectors of the same dimension is extracted for both VIS and IR images. A fully connected layer accepting 2048-dimensional feature vectors and outputting scores for number of individuals in the training set is defined for baseline. If the distance based score layer is used, the number of class centers are chosen to be equal to the number of individuals in the training set. Network architecture with the distance based score layer is given in Figure 2.

3.3.1 Identification Loss

In [37], it is emphasized that the identification loss has higher discriminative ability than the verification loss since the verification loss assigns weak labels and adjusts distances between pairs mostly. With the identification loss, the network focuses on small details to identify each person whose appearances might be similar and difficult to distinguish. Therefore, the network with identification loss learns to distinguish not only easy samples but also difficult ones. It is also considered that the use of a distance metric in the identification loss helps distances of positive pairs get closer and negative pairs get apart. IR and VIS images belonging to the same person are mapped to the same class. The loss is calculated based on the correct and incorrect person classifications using the softmax cross-entropy loss function.

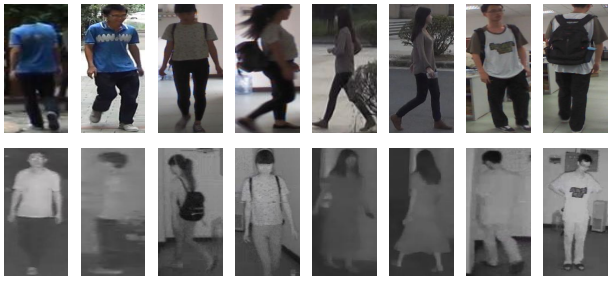

Figure 3. Example images of four different subjects from SYSU-MM01 dataset [27]. VIS images are given in the top row and NIR images are given in the bottom row.

When identification loss is combined with the distance based score layer, correct classification is achieved only if the distance between a sample and its corresponding center point are closer than the distance between the sample and other center points. Therefore, the distance based score layer with the identification loss behaves similarly to contrastive and triplet loss functions and can achieve lower distance for positive samples than negative samples in both intra-modality and inter-modality settings.

3.3.2 Feature Extraction and Ranking

In the ranking stage, all test images are forward-passed through the network and their corresponding feature vectors are stored. Then, the feature vectors of gallery images are compared with the feature vector of the query image. The aim of the ranking stage is to find images closest to the query image at highest ranks. Therefore, the gallery images are sorted for each query image. All distance metrics are used during ranking stage and their corresponding performance is found for each setting.

4. Experiments

This section presents experimental evaluation of the proposed architecture. The experiments are carried out on two cross-modality datasets and results are compared with the state-of-the-art works in cross-modality person re-identification.

4.1. Datasets

SYSU-MM01 cross-modality person re-identification dataset [27] is the first dataset used in the experiments. This dataset contains visible and NIR images of 491 identities taken from 6 cameras. A total of 287,628 VIS images and 15,792 NIR images are found in the dataset. The dataset contains images shot in indoor and outdoor environments with dark and bright conditions. Out of six cameras, four cameras work in VIS bands and two cameras work in NIR bands of the electromagnetic spectrum. Example images from SYSU-MM01 dataset are given in Figure 3. Images in the dataset are resized via stretching or shrinking to the size of 224x224 to make them compatible for feeding to our Resnet based network. Training and test splits are constructed as in the original work [27] and images of 296 different individuals are used in the training.

Beta	r1	r10	r20	mAP
0.10	25.02	67.39	81.47	25.98
0.30	25.71	70.86	84.00	27.16
0.50	26.31	73.65	86.57	28.46
0.70	**28.84**	**74.37**	**87.14**	**30.55**
0.90	26.43	69.92	84.12	28.33

Table 1. Impact of β on performance of the distance based score layer with L2 distance in *all-search* / *single-shot* test settings

RegDB dataset [20] is the second dataset used in the experiments. RegDB dataset contains images taken by VIS and thermal cameras mounted on the same plate. The dataset contains images of 412 persons. It includes 10 color and 10 thermal images for each person. A total of 4120 thermal images and 4120 VIS images are contained in the dataset. Images are resized to 224x224 to feed them into our ResNet model.

4.2. Implementation Details

Stochastic Gradient Descent algorithm is used for parameter updates in the training stage. Learning rate of 0.01 and momentum of 0.9 are used as hyperparameters of SGD. Batch size of 32 is employed throughout the experiments. Batches are randomly constructed from either IR (NIR or thermal) or VIS images and fed into the network. Initialization of center points in the distance based score layer is performed with random values.

4.3. Evaluation Protocol

The evaluation protocol follows the same standard procedure given in [27] for SYSU-MM01 dataset. Cumulative Match Score (CMS), which is giving the percentage of successful occurences at a given rank, and Mean Average Precision (mAP) [36] metrics are used for evaluation. Performance evaluation of the proposed method is performed using the code provided by the dataset owner [27]. NIR images are used as query images and the gallery set is constructed from VIS images. Test images are divided into 10 splits and average of mAP and CMC scores are calculated. In addition, performance scores for all other settings including *all-search*, *indoor-search*, *single-shot*, *multi-shot* are calculated.

The same procedure of [31] is used during evaluation on RegDB dataset. Training and ranking operations are performed for each trial (a total of 10 trials) and their average is calculated for statistical stability. Visible to thermal query

Training Style		Ranking Metric											
Center Update Method	Training Metric	L1				L2/L2-Squared				L∞			
		r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
SGD	Fully Connected	28.61	73.09	85.88	29.20	28.30	72.68	85.00	28.69	19.36	56.80	71.69	18.86
SGD	Dot Prod.	27.46	71.35	84.29	28.06	27.13	70.52	83.33	27.68	17.63	53.97	69.03	17.48
SGD	L1	1.60	13.03	24.74	3.49	1.58	12.75	23.87	3.43	1.35	12.34	22.80	3.28
SGD	L2	12.15	50.58	65.78	16.52	12.02	49.97	65.06	16.33	11.70	46.48	60.85	15.09
SGD	L2-Squared	**23.88**	**67.52**	**83.32**	**24.30**	***23.68***	***67.10***	***82.35***	***24.07***	14.61	50.32	65.89	14.86
SGD	L∞	5.44	31.66	48.45	8.50	5.60	32.45	49.19	8.77	4.94	30.25	47.21	8.21
Exp.Run.Avg.	L1	5.01	29.07	46.58	7.28	4.83	28.11	45.16	7.07	2.92	19.72	34.22	4.89
Exp.Run.Avg.	L2	**29.05**	**74.71**	**87.16**	**30.94**	***28.84***	***74.37***	***87.14***	***30.55***	22.97	64.44	78.81	22.69
Exp.Run.Avg.	L2-Squared	10.66	44.51	62.60	12.78	10.87	45.00	62.91	12.99	6.34	32.19	49.00	8.04
Exp.Run.Avg.	L∞	21.24	63.54	78.89	23.54	19.26	62.05	78.41	22.89	17.09	57.67	73.83	19.98
Median	L1	5.09	29.80	47.00	7.29	4.62	28.56	45.57	6.91	2.82	19.70	33.81	4.81
Median	L2	**27.49**	**71.95**	**85.88**	**29.71**	***26.92***	***71.34***	***85.65***	***29.19***	20.49	61.39	76.44	21.27
Median	L2-Squared	9.61	43.08	60.75	11.81	9.45	42.43	59.91	11.57	5.53	29.50	45.49	7.28
Median	L∞	19.07	62.69	79.71	22.03	16.24	58.04	75.17	20.04	14.08	53.85	70.61	17.41
Mid-range	L1	3.65	23.44	39.39	5.67	3.22	22.30	37.91	5.30	1.90	14.72	27.37	3.71
Mid-range	L2	**30.37**	**73.28**	**86.06**	**31.78**	***29.92***	***72.84***	***85.89***	***31.37***	24.18	64.93	78.98	24.09
Mid-range	L2-Squared	15.24	54.59	72.26	16.94	15.12	54.47	72.37	16.85	9.78	42.98	60.22	11.22
Mid-range	L∞	21.93	64.13	78.34	24.38	20.63	64.38	79.36	24.10	17.89	59.80	75.19	20.72

Table 2. Comparison of the training methods with different update schemes and different training/ranking metrics in *all-search* / *single-shot* test settings on SYSU-MM01 dataset

setting results are presented in the following sections.

4.4. Experimental Results

Experiments are carried out to investigate different configurations and parameters of the distance based score layer on SYSU-MM01 [27] dataset. Then, performance of the proposed method is compared with the state-of-the-art methods on both SYSU-MM01 and RegDB [20] cross-modality person re-identification datasets.

When exponential running average is used for updating center points, parameter β adjusts the update rate of center points and has an effect on network training. Different values of parameter β are tested for the case that L2 distance is used in the training and ranking stages. The effect of parameter β on the performance is given in Table 1 for *all-search* / *single-shot* test settings. Parameter β should be selected carefully for obtaining better results. Large β results in slow update rate of center points while small β leads to fast changing center points. As seen in Table 1, the method performs better if β is around 0.70 on SYSU-MM01 dataset. Thus, in the rest of the experiments, β is set to 0.70 for exponential running average update method.

In Table 2, performance of using L1, L2, L2-squared, L∞ distance metrics, and negative of dot product in the distance based score layer during training stage are listed. In addition, the results of training the network with fully connected layer are also given. The same hyperparameters and training methods are used during comparison. Ranking results are given using all distance metrics for each combination of training settings. It is observed that L2 distance performs the best among other configurations and distance metrics during training. It is also seen that L1 distance metric performs best in the ranking stage among other distance metrics. The widely used method of training with the fully connected layer performs as the second best. Although L1 distance is succesful in ranking stage, using L1 distance during training leads to very poor performance and under-trained network. Exponential running average, median and mid-range center update methods are succesful when L2 and L∞ distance metrics are used. However, it is observed during development that exponential running average is more stable than median and mid-range update schemes and generally leads to higher performance. The results obtained using negative of dot product in the distance based score layer is very close to the results of the fully connected layer. The operations in negative of dot product configuration are similar to operations of the fully connected layer but with a different layer structure. This result also shows that the distance based score layer works as intended.

Four different state-of-the-art methods that investigate cross-modality person re-identification on SYSU-MM01 dataset are considered for comparison with the proposed method. Deep zero padding one-stream architecture by Wu *et al*. [27] obtains superior scores than the algorithms based on hand-crafted features. Hierarchical discriminative learning architecture of Ye *et al*. [31] is another study that in-

Method	All-search								Indoor-search							
	Single-shot				Multi-shot				Single-shot				Multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
Asymmetric FC [27]	9.30	43.26	60.38	10.82	13.06	52.11	69.52	6.68	14.59	57.94	78.68	20.33	20.09	69.37	85.80	13.04
One-stream [27]	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
Two-stream [27]	11.65	47.99	65.50	12.85	16.33	58.35	74.46	8.03	15.60	61.18	81.02	21.49	22.49	72.22	88.61	13.92
Zero-Padding [27]	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
TONE + HCML [31]	14.32	53.16	69.17	16.16	–	–	–	–	–	–	–	–	–	–	–	–
BCTR [32]	16.12	54.90	71.47	19.15	–	–	–	–	–	–	–	–	–	–	–	–
BDTR [32]	17.01	55.43	71.96	19.66	–	–	–	–	–	–	–	–	–	–	–	–
cmGAN [5]	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
Ours (Dist. based)	**29.05**	**74.71**	**87.16**	**30.94**	**35.40**	**81.02**	**91.85**	**24.12**	**32.74**	**82.40**	**93.35**	**44.26**	**40.41**	**86.83**	**96.27**	**33.93**

Table 3. Comparison with the state-of-the-art methods on SYSU-MM01 cross-modality person re-identification dataset. The distance based method indicates the network with the distance based score layer whose variables are updated with exponential running average.

Method	r1	r10	r20	mAP
One-stream [27]	13.11	32.98	42.51	14.02
Two-stream [27]	12.43	30.36	40.96	13.42
Zero-Padding [27]	17.75	34.21	44.35	18.90
TONE + HCML [31]	24.44	47.53	56.78	20.80
BCTR [32]	32.67	57.64	66.58	30.99
BDTR [32]	33.47	58.42	67.52	31.83
Ours (Dist. based)	**38.64**	**60.18**	**69.81**	**38.08**

Table 4. Comparison with the state-of-the-art methods on RegDB cross-modality person re-identification dataset in visible to thermal settings.The distance based method indicates the network with the distance based score layer whose variables are updated with exponential running average.

cludes feature learning and metric learning stages. Dual path network architecture of Ye *et al*. [32], which contains identification loss and dual-constrained top-ranking loss, and triple-stream network architecture of Dai *et al*. [5], which contains identification loss, triplet constraints, and modality classifier trained with GAN are included in the comparison. The results of the distance based score layer are compared with these studies in Table 3. Results of the proposed architecture are given in the last row of the comparison table. The exponential running average method with L2 training and L1 ranking is employed in this experiment.

As observed in Table 3, the comparative results include rank-1 (r1), rank-10 (r10), and rank-20 (r20) accuracies of CMC and mAP scores in different test settings. The proposed one-stream architecture with the distance based score layer outperforms the state-of-the-art methods in all categories on SYSU-MM01 dataset. In addition, the proposed distance based score layer improves performance over the fully connected layer in the case that L2 distance metric is used in the training stage.

Comparisons with the state of the art methods on RegDB dataset is given in Table 4. Similar to the comparison done on the first dataset, the exponential running average method with L2 training and L1 ranking is used in this experiment. As update method, exponential running average is employed. The proposed architecture improves both CMC and mAP scores compared to the state of the art methods.

The results demonstrate that the proposed distance based score layer can be a good alternative to the fully connected layer in classification and/or ranking tasks. As we have shown on cross-modality person re-identification task, the distance based score layer can perform better than the fully connected layer, when suitable distance metric is used for the application.

5. Conclusion

An one-stream network architecture and the distance based score layer are proposed for person re-identification task in cross-modality images. The proposed approach aims to use distance metrics in both the training and ranking stages in order to improve performance. The distance based score layer calculates distances between a feature vector and center points of each class in the training stage. Center points are updated as the network gets trained and class centers are optimized to find the identity corresponding to the given feature vector using a distance metric. Thus, the distance based score layer enables to use a distance metric in the training of CNNs and uncovers the opportunity to include the desired distance metric in deep learning applications. In the future work, the distance based score layer can be studied with different structures and distance metrics. Since this layer can be easily adapted to other classification and identification problems, there seems to be great potential in the future deep learning applications.

6. Acknowledgement

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant No. 114G028.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916, 2015.

[2] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person re-identification. *arXiv preprint arXiv:1801.05339*, 2018.

[3] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016.

[4] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[5] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, pages 677–683, 2018.

[6] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

[9] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[13] Felix Juefei-Xu, Dipan K Pal, and Marios Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 141–150, 2015.

[14] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[15] José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6628–6637, 2017.

[16] Stan Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013.

[17] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013.

[18] Xiaoxiang Liu, Lingxiao Song, Xiang Wu, and Tieniu Tan. Transferring deep representation for nir-vis heterogeneous face recognition. In *2016 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2016.

[19] Pattabhi Ramaiah Nalla and Ajay Kumar. Toward more accurate iris recognition using cross-spectral matching. *IEEE transactions on Image processing*, 26(1):208–221, 2017.

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

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[22] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3960–3969, 2017.

[23] Evgeniya Ustinova, Yaroslav Ganin, and Victor Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[24] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016.

[25] Ritesh Vyas, Tirupathiraju Kanumuri, and Gyanendra Sheoran. Cross spectral iris recognition for surveillance based applications. *Multimedia Tools and Applications*, pages 1–19, 2018.

[26] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 274–282. ACM, 2018.

[27] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE Interna-*

tional Conference on Computer Vision, pages 5380–5389, 2017.

[28] Lin Wu, Chunhua Shen, and Anton van den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65:238–250, 2017.

[29] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016.

[30] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *European conference on computer vision*, pages 536–551. Springer, 2014.

[31] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, pages 1092–1099, 2018.

[33] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014.

[34] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[35] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017.

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

[37] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018.