Cross-Visual Attention Fusion Network with Dual-Constrained Marginal-Ranking for Visible-Infrared Person Re-Identification

Fang Su¹,³, Meibin Qi¹,³*, Cuiqun Chen¹,³, Tan Bo¹,³ and Jianguo Jiang¹,³

¹DSP Laboratory, School of Computer and Information, Hefei University of Technology, Hefei, Anhui, 230009, China

ªfs@mail.hfut.edu.cn; *qimeibin@hfut.edu.cn; ªchencuiqun_hfut@163.com;
³tanbo.hfut@foxmail.com; ªjgjiang@hfut.edu.cn

Abstract. Visible-Infrared Person re-identification(VI-REID) is extremely important for night-time surveillance applications. It is a challenging problem due to large cross-modality discrepancies and intra-modality variations caused by different illuminations, human poses, viewpoints, etc. In this paper, we propose a cross visual attention fusion dual-path neural network with dual-constrained marginal ranking(DCAF) to solve the problem. First, we utilize cross-visual attention to learn discriminative feature of high-level semantic information in their respective modals. Second, in order to establish the relationship between modals, we fuse attentional weight of two modals and add it into backpropagation to obtain those regions that are distinctive for classification. Third, a dual-constrained marginal-ranking loss is introduced to narrow the gap between different networks and to learn strongly the similarities between two modals. Extensive experiments demonstrate that the proposed approach effectively improves the performance of VI-REID task and remarkably outperforms the state-of-the-art methods.

1. Introduction
Person re-identification aims at connecting the human body across cameras and temporal periods in a non-overlapping perspective. From the last decade¹,²,³,⁴ to recent studies⁵,⁶,⁷, re-id has received lots of attention. However, most RGB-RGB person re-identification methods can only solve the problem of recognition in sufficient light, while at night or in dark light, the visible camera is almost useless. Criminals, or people with an attempt, tend to move at night, when it is obviously impossible to solve the problem just depending on the visible camera to capture images. Thus, the re-identification results may be not robust due to the limit of single model.

In the field of RGB-IR, this paper studies the pedestrian re-identification under the condition of cross-modality, and mainly solves the problem of matching images in cross-modality. In other words, query set is an image in one modal, such as infrared image, while gallery set is an image in another modal, such as RGB image. This research is of great significance on REID. Cross-modality person re-identification is a new problem that has emerged in the last two years, so few academic research results can be found in this field.⁸,⁹ have proposed some methods for addressing VI-REID, the identification accuracy is not high. Inspired by several recent works¹⁰,¹¹, we propose a cross-visual attention to solve the RGB-IR Person re-identification( RGB-IR) problem.
The main contributions can be summarized as follows: 1) We introduce dual-constrained marginal-ranking constraint by affine transformation to narrow the gap between different networks. Dual-constrained marginal-ranking is a powerful feature representation that combines two constraints to strengthen the cross-modality and intra-modality similarity. 2) We propose a cross-visual attention network with GRU, which is the first time that sequential model with GRU is applied to the VI-REID problem. We embed the features of attention into the same space by fusion, then we add features into backpropagation to get the distinguishing regions. 3) We demonstrate experimentally the superior performance of our method on SYSU-MM01[8] and RegDB[12]. For SYSU-MM01 dataset, 35.88% mAP and 34.66% top-1 accuracy are obtained under all-search modal and 49.88% mAP and 39.23% top-1 accuracy are obtained under indoor-search modal, while on RegDB dataset, obtaining 41.66% mAP and 40.42% top-1 accuracy under all search modal.

2. The Proposed Method

2.1. Framework

This paper proposes a framework of cross-visual attention fusion dual-path neural network with dual-constraints marginal-ranking constraints (DCAF) for VI-REID. Figure 1 shows the overall framework of our proposed approach. This framework has learned the high-level feature representation and measurement in an end-to-end way. To be specific, we have dual stream path, the backbone of each path add respectively a parallel attention. Based on the attention features gained above, we combine separately the visible and infrared attention features (att_v, att_t) with the global features (f_v, f_t) by addition to obtain more comprehensive information. Then, two modals are embedded into the common space by affine transformation, we use cross-modality marginal-ranking constraint and euclidean constraint to narrow the gap between dual networks. Specifically, we map \( x_1 \) and \( x_2 \), vi_att and th_att to a common space by sharing the full connection layer, thus reducing the difference between infrared and visible modals. It is worth noting that the weight feature is extracted because attention is measured the value of information, so as to determine the search scope. The greater the weight distribute, the higher the value get. We embed the attention weight into the common space and add it into backpropagation to obtain the regions that are differentiated for classification.
Figure 1. The proposed dual-path end-to-end learning framework by recurrent model for VI-REID. To be specific, \( f_v \) and \( f_t \) represent respectively global feature of each modal. Then, We add the global feature after average pooling to the local feature after recurrent model. We use the recurrent neural network with GRU to model the high-level features to get more comprehensive feature. The specific recurrent model is shown in the Figure 2(b). A feature vector is obtained by CNN, and the feature slices \( C \) are obtained by attention mechanism, then used it as input of the GRU. \( l_1 \) is a given initial value for us. The next location probability \( l_{i+1} \) is predicted by GRU propagation. \( y_i \) represents class label. \( \varphi_1, \varphi_2, \varphi_3, \varphi_4 \) are projected respectively into a common space to strengthen the similarity between two modals.

2.2. GRU NETWORK AND THE ATTENTION MECHANISM

In our attention model, we use the GRU network to generate new category labels and new hidden layer states through the upper and lower picture vectors of each time step and the characteristics of the previous hidden layer states. We use the GRU in the following way[13]. At the time step \( t \), the GRU uses a masked CNN feature map \( C \) and the previously hidden state \( h_{t-1} \) as input, which is also illustrated in Figure 2. Note that the graph is predicted from the previously hidden state \( h_{t-1} \) using the parameters of learning \( W_{i,h} \). Then, an input feature map with a predicted attention map \( l_{i-1} \) mask size of \( K*K*D \) is used to obtain a filtered feature map \( X_t \) with only the participating areas retained. The formula is as follows:

\[
\begin{align*}
    z_t &= \sigma(M_z[h_{t-1}, C_t]) + b_z \\
    r_t &= \sigma(M_r[h_{t-1}, C_t]) + b_r \\
    f_t &= \tanh(M_f[h_{t-1}, C_t]) + b_f \\
    h_t &= (1 - z_t) + z_t \odot h_{t-1}
\end{align*}
\]

In the formula, \( z_t, r_t, f_t, h_t \) are the update gate, reset gate, memory unit and hidden state of the GRU[13] respectively. \( M_z, M_r, M_f \) and \( b_z, b_r, b_f \) represent learnable weight parameters within the gates and and \( \odot \) is the dot product. And then \( C_t \) represents the GRU input at time-step \( t \). The input of the image after FCN network is a feature vector of \( 32*2048 \) dimension. This attention divides the feature of \( 2048 \) into four \( 512 \)-dimension patches. The experiment proves that setting time-steps as \( 4 \) has the best effect, and different sizes of patch can be set according to specific needs.
2.3. Feature fusion and embedding

Multi-modality information fusion can obtain more comprehensive features and improve the robustness of the modal. In terms of feature fusion, concatenation and addition are two fusion methods adopted in this paper. It is worth noting that concatenation operates on the number of channels, which can increase the number of channels, while addition operates on the number of feature graphs, which does not change the number of channels in the image. We use affine transformation to embedding visible- and thermal- modality into common space, so as to get the commonness of two modals. In this paper, feature fusion and embedding are used in the following aspects. First, the high-level semantic information is obtained with the visible and infrared images going through backbone. On this basis, we introduce an average pooling layer to obtain the global information of the images, then the global information and the local information of attention features are added respectively to obtain more comprehensive features. Second, the purpose of attention redistribute the weight of features. Without the Attention mechanism, we can assume that every feature of input has the same effect on every feature of output, while attention is further distinguishing the different importance of features and assigning different weights of attention, which emphasize the importance of different features. Therefore, the attention weight$(vi\_att, th\_att)$ and feature map$(x_1, x_2)$ are concatenated respectively to get shared features. Finally, we use affine transformation to embed the attention weight$(vi\_att, th\_att)$ and feature map$(x_1, x_2)$ into common space, so as to get the similarity between two modals.

2.4. LOSS FUNCTIONS

In order to give full play to the recognition ability of learning representation in the network structure, we employ $L_{\text{softmax}}$ loss for classification, then cross modality marginal-ranking constraint and euclidean constraint for metric learning as the loss functions in training phases, which get good effect with dual constraint. Then we describe the losses we used in detail.

$$L_{\text{Triplet}} = - \sum_{i=1}^{p} \sum_{a=1}^{K} (\alpha + \max_{p=1..k} \|x_a^{(i)} - x_p^{(i)}\|_2 - \min_{n=1..k} \|x_a^{(i)} - x_n^{(j)}\|_2}_2, j \neq i, p)$$  (5)
The loss function is formulated as follows:

\[ L_{\text{classification}} = \frac{1}{N} \sum_{i} - \log(\frac{e^{y_i \alpha}}{\sum_j e^{y_j \alpha}}) \] (6)

We use cross-modality marginal ranking constraint, which penalizes the modal according to the hardest negative examples. Inspired by the loss function commonly used in structured prediction [14,15], we focus on the hard negatives points of training, i.e., the negatives points closest to each training query. This is particularly relevant for cross-modality retrieval, as it is the most detrimental factor in determining the success or failure of a measure. In the following form, \( \phi_1 \) and \( \phi_2 \) means parameters of visible path. \( \phi_3 \) and \( \phi_4 \) means parameters of thermal path. Concretely, the loss \( L_{\text{CMR1}} \) indicates that \( \phi_1 \) and \( \phi_2 \) are constrained, the loss \( L_{\text{CMR2}} \) indicates that \( \phi_3 \) and \( \phi_4 \) are constrained. The loss function is formulated as follows:

\[
L_{\text{CMR1}} = \max(0, \alpha + s(\phi_1, \phi_2) - s(\phi_1, \phi_2)) + \max(0, \alpha + s(\phi_1, \phi_2) - s(\phi_1, \phi_2)) \] (7)

\[
L_{\text{CMR2}} = \max(0, \alpha + s(\phi_3, \phi_4) - s(\phi_3, \phi_4)) + \max(0, \alpha + s(\phi_3, \phi_4) - s(\phi_3, \phi_4)) \] (8)

Firstly, we use dual-constrained marginal-ranking to make the distance between the same class of features as small as possible and the distance between different class as large as possible. Secondly, L Softmax has the following advantages of being able to adjust the margin of expectation and avoid overfitting, as well as being optimized by the typical stochastic gradient descent method. Finally, cross-modality marginal-ranking constraints and triplet loss and classification loss are combined as the final loss.

\[
L = L_{\text{Triplet}} + L_{\text{CMR1}} + L_{\text{CMR2}} + L_{\text{Classification}} \] (9)

3. Experiment Result

3.1. Datasets

Datasets: Two datasets SYSU-MM01 [8] and RegDB [12] are used for experimental evaluations. The SYSU-MM01 dataset is divided into training set and test set. The training set contains 395 pedestrians, 22,258 RGB images and 11,909 infrared images, and the test set contains 96 pedestrians. There are two search modes: all search mode and indoor search mode. RegDB [12] contains 412 persons collected by 2 cameras, one visible camera and one thermal camera. For each person, 20 different images are captured by both visible and thermal cameras.

3.2. Implementation

First, this article chooses the single-shot model to test it and takes the Stochastic Gradient Descent (SGD) to fine-tune network. Second, the size of the embedding fully connected layer is seted as 1024 and the 265 batch size is seted as 16, the initial learning rate is seted as 0.001, the learning rate in the 30th epoch reduced to 0.0001 times. Third, the pre-defined margin for the triplet loss is seted as 1.2. Fourth, the entire network carries on the 60 epochs.

| Method          | All-search   | Indoor-search |  
|-----------------|--------------|---------------|
|                 | Rank-1       | Rank-20       |
|                 | Rank-10      | Rank-10       |
|                 | mAP          | Rank-20       |
|                 | mAP          |               |
| LOMO[16]        | 1.75         | 26.63         |
| HOG[17]         | 2.76         | 4.24          |
| One-stream[18]  | 12.04        | 13.67         |
| Two-stream[18]  | 11.65        | 12.85         |
| Zero-Padding[18]| 14.80        | 15.95         |
| TONE[19]        | 12.52        | 68.60         |

Table 1. Evaluation on SYSU-MM01. Re-identification rates(%) at rank r and mAP(%)
Table 2. Evaluation on RegDB. Re-identification rates(%) at rank r and mAP(%) 

| Method       | Rank-1 | Rank-10 | Rank-20 | mAP  |
|--------------|--------|---------|---------|------|
| LOMO[16]     | 0.85   | 2.47    | 4.10    | 2.28 |
| HOG[17]      | 13.49  | 33.72   | 43.66   | 10.31|
| One-stream[18]|13.11  | 32.98   | 42.51   | 14.02|
| Two-stream[18]|12.43  | 30.36   | 49.96   | 13.42|
| GSM[23]      | 17.28  | 34.47   | 45.26   | 15.06|
| Zero-Padding[18]|17.75 | 34.21   | 44.25   | 18.90|
| TONE[19]     | 16.87  | 34.03   | 44.10   | 14.92|
| HCNL[19]     | 24.44  | 47.53   | 56.78   | 20.80|
| BDTR[20]     | 33.47  | 58.42   | 67.52   | 31.83|
| DCAF(Ours)   | 40.42  | 62.56   | 72.10   | 41.66|

3.3. Performance Comparison
Several used commonly methods are evaluated on SYSU-MM01 dataset in Table 1 and on RegDB dataset in Table 2 to verify the effectiveness of our method. These methods include some feature learning methods such as LOMO[16], HOG[17], One-stream and Two-stream[18]. In addition, we compare our method with most of the related methods for VI-REID. These methods include Zero-Padding[18], TONE[19], HCNL[19], BDTR[20], besides, two GAN methods are contained : cmGAN[21], D³RL[22].

3.4. Effect of Major Components
In order to comprehensively evaluate the effectiveness of this method, we conducted ablation experiments for different component settings in single query mode on SYSU-MM01 dataset.

Table 3. Analysis of each component on the SYSU-MM01 dataset

| SYSU-MM01       | All-search | Indoor-search |
|-----------------|------------|---------------|
|                 | Rank-1     | Rank-10       | Rank-20 | mAP  | Rank-1 | Rank-10 | Rank-20 | mAP  |
| baseline        | 23.01      | 65.53         | 81.51   | 24.98 | 26.99  | 75.09   | 90.04   | 38.42|
| only attention  | 32.79      | 75.49         | 88.72   | 33.39 | 35.97  | 86.56   | 95.61   | 46.34|
| No marginal-ranking | 28.40  | 71.00         | 83.88   | 29.10 | 34.35  | 83.68   | 90.99   | 46.89|
| No triplet loss | 30.82      | 74.52         | 88.35   | 32.97 | 36.21  | 85.35   | 92.04   | 47.91|
| Full model      | 34.66      | 77.57         | 89.90   | 35.88 | 39.23  | 89.68   | 95.61   | 49.88|

As can be seen from the table 3, we could observe that GRU attention achieves rank-1 accuracy about 32.79%, while mAP about 33.39%. Although the improvement is not significant after fusion, the combination still improves the performance. Meanwhile, the rank-1 matching rate is about 28.40% for the triplet loss while the mAP is about 29.10% on the SYSU-MM01 dataset. After integrating the marginal-ranking loss, our final full model could achieve rank-1 = 34.66%, and mAP = 35.88%.

4. Conclusion
This paper puts forward a framework of end-to-end dual-path neural network for cross-visual attention fusion with dual-constrained in terms of the problems of VI-REID. In order to model high-level semantics and correlate the information between modals, we use the cross-visual attention model of
the recurrent neural network with GRU to process sequence information. Based on the obtained attention features, thermal- and visible- attention features are fused and directly added into the backpropagation for classification training, which obtain those discrimination regions. Then the attention features are projected into the common space by affine transformation, and the cross-modality marginal-ranking constraint and euclidean constraint are used to narrow the gap between the two modals. Experimental results show that our method has advanced performance in the field of VI-REID.

Acknowledgments
This work is greatly supported by Natural Science Foundation of China (No. 61771180 and No. 61876056). Thank you very much for encouragement.

References
[1] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person reidentification," in CVPR, 2016, pp. 1268-1277.
[2] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in CVPR, 2013, pp. 3318-3325.
[3] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in CVPR, 2012, pp. 2288-2295.
[4] Wu J, Jiang J, Qi M, et al. Independent metric learning with aligned multi part features for video-based person re-identification[J]. Multimedia Tools and Applications, 2019: 1-19.
[5] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," arXiv preprint arXiv:1701.07717, vol.3, 2017.
[6] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," arXiv preprint arXiv:1711.10295, 2017.
[7] G.Wang, J.Lai, and X.Xie,"P2sn: Can an image match a video for person re-identification in an end-to-end way?"IEEE TCSVT, vol.99, pp.1-1,2017.
[8] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "Rgb-infrared cross modality person re-identification," in ICCV, 2017, pp. 5380-5389.
[9] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," 2018.
[10] K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., Jun. 2015, pp. 2048-2057.
[11] D. Bahdanau, K. Cho, and Y. Bengio. (Sep. 2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available:https://arxiv.org/abs/1409.0473
[12] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras", in proc. Sensors., 2017.
[13] Wang,Liwei,Li,Yin, and Lazebnik,Svetlana. Learning two-branch neural networks for image-text matching tasks. arXiv preprint arXiv:1704.03470, 2017.
[14] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),32(9):1627-1645, 2010.
[15] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun.Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research (JMLR), 6(Sep):1453-1484, 2005.
[16] Liao S C, Hu Y, Zhu X Y, et al. Person re-identification by local maximal occurrence representation and metric learning[C]// Proceedings of the 2015 440 IEEE Conference on Computer Vision and Pattern Recognition. Boston,MA, USA: IEEE, 2015:2197-2206.
[17] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in CVPR, vol. 1, 2005, pp. 886-893.
[18] Wu, A., Zheng, W. S., Yu, H. X., Gong, S. and Lai, J. "RGB-infrared cross-modality person re-identification". In ICCV. 2017.
[19] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in CVPR, vol. 1, 2018, p. 2.
[20] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person reidentification via dual-constrained top-ranking," in IJCAI, 2018.
[21] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person reidentification with generative adversarial training", in proc. International Joint Conferences on Artificial Intelligence., 2018.
[22] Z. Wang, Z. Wang, Y. Zheng, Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification", in proc. IEEE Conf. Comput. Vis. Pattern Recognition. 2019.
[23] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, “Cross-domain visual matching via generalized similarity measure and feature learning,” in TPAMI, 2017, pp. 1089–1102