Songklanakarin J. Sci. Technol.
46 (5), 466–475, Sep. – Oct. 2024

http://www.sjst.psu.ac.th

Original Article

A study on Chinese language Cross-Modal pedestrian image information retrieval

Yan Xie, and Jian Qu*

Faculty of Engineering and Technology, Panyapiwat Institute of Management, Pak Kret, Nonthaburi, 11120 Thailand

Received: 30 October 2023; Revised: 1 August 2024; Accepted: 7 August 2024

Abstract

This research aims to achieve cross-modal Chinese pedestrian information retrieval (CMCPIR). Most PIR tasks focus on English due to data availability, as to the best of our knowledge, there is no available Open access Chinese dataset or PIR. Therefore, we built a Chinese dataset (CUHK-PEDES-CHINESE) using English CUHK-PEDES dataset. We constructed a baseline CPIR model (HCMacBRes50) by modifying a top performing English PIR model. We have experimented and modified a set of deep learning models based on image model and Chinese language text model to construct a novel RoberResX4D CMCPIR model. Furthermore, we designed a novel MSWCMPME loss function to enhance the performance of our CMCPIR model by optimizing the fusion strategy for the outputs of image-text model. The RoberResX4D model has the same PIR performance and is more robust than the English model of SRCF. The RoberResX4D model effectively combines information from different modalities.

Keywords: Chinese pedestrian information retrieval, convolutional neural network, loss function, Cross-Modal, deep learning

1. Introduction

Nowadays, CCTV cameras are installed in most public places to acquire pedestrian images. However, pedestrian images acquired from actual surveillance videos are often of low-quality. The main reasons for the low quality of pedestrian images are low resolution, uneven lighting, viewing angle problems and complex backgrounds. The low resolution of an image can make it difficult to distinguish details of the pedestrian such as facial features and clothing textures. Viewing angle problems can lead to distorted pedestrian images and obscured feature information. Complex backgrounds can make it more difficult to clearly separate pedestrian images. Low-quality images make it difficult to extract highly representative facial features for identification. Low-quality images should not be used directly for pedestrian retrieval. Therefore, existing research has shown that deep learning can be used to handle pedestrian detection methods for low quality images. This technique of using computer vision to determine the presence or absence of specific pedestrians in a dataset is called Pedestrian Information Retrieval (PIR). PIR research focuses on how to learn and mine the matching relationship between images and text (Li, Xiao, Li, Yang, & Wang, 2017). The PIR is achieved by learning the relationship between image and text. Textual descriptions usually accompany the images. These descriptions may include details of the appearance, clothing, accessories and other distinguishing features of the pedestrian. Generally speaking, PIR problems can be solved in two different methods: the traditional PIR methods and the PIR feature extraction methods.

Cao, Araújo, and Sim (2020) trained a DELG model to implement PIR for images. However, the use of image-based search for image methods greatly limits the degrees of freedom of PIR. Therefore, existing research (Li, Xiao, Li, Yang, & Wang, 2017; Zhu *et al.*, 2021; Yang *et al.*, 2021) proposes the use of natural language descriptions to achieve PIR. Additionally, Chen *et al.* (2021) demonstrated that the cross-modal PIR model performs better than the uni-modal PIR model in the existing studies. Therefore, we propose to use natural language-based cross-modal PIR in our research.

*Corresponding author
Email address: jianqu@pim.ac.th

Our research focuses on the appearance features of pedestrians. Appearance features of pedestrians mainly include the length of hair/clothes/pants and the color of clothes/pants/shoes. Cao *et al.* (2020) used manually designed feature operators for feature representation in their study. However, the study of Li *et al.* (2017) showed that the features computed by simple machine learning algorithms have difficulty in meeting the practical demands of complex scenarios and large data volumes. Therefore, Ding, Ding, Shao, and Tao (2021) proposed to use deep neural networks to enhance feature extraction in their study. Existing studies have shown that feature extraction using deep neural network models improves accuracy well compared to traditional feature methods (Chen *et al.*, 2021; Farooq, Awais, Kittler, & Khalid, 2021; Suo *et al.*, 2022). The deep learning model overcomes the problem of weak learning representation and applicability to massive data in traditional feature extraction methods (Chandrakala & Sim, 2021; Li & Qu, 2022). Therefore, we proposed to use deep learning models for feature extraction in the PIR feature extraction module. Furthermore, we found that existing studies have not yet used Chinese corpus to train PIR. English and Chinese are the two most frequently used languages in the world. Most PIR tasks focus on English due to data availability. We have not found any papers on Google Scholar that use Chinese language in the PIR task. This is most likely due to the unavailability of the dataset. Therefore, we proposed to train cross-modal PIR based on Chinese natural language descriptions. We called PIR based on Chinese natural language descriptions "Chinese Pedestrian Information Retrieval" (CPIR). In summary, we train a CPIR model facing the challenges of effectively fusing data from different modalities of text and images, the lack of Chinese datasets for pedestrian information retrieval, and the difficulty of labeling Chinese data.

We split CPIR into two components: the dataset construction component and the PIR model optimization component. In the dataset construction component, we found that the CUHK-PEDES dataset is widely validated for PIR tasks in existing studies (Chen *et al*., 2021; Suo *et al*., 2022). Meanwhile, we consider the use of real-time datasets to demonstrate the efficiency and effectiveness of the method. Experiments by Li *et al.* (2017) show that data outside the dataset used to validate pedestrian information retrieval is very ineffective. In terms of pedestrian information retrieval performance, the robustness of models trained on a small amount of data is extremely poor. The research of Yufeng and Guoxiu (2023) shows that models trained on a small amount of data must be sufficiently small or noiseless during the retrieval process. Furthermore, in the Thailand Super AI project, we collected our own dataset of pedestrian information retrieval. There were about thirty people, and we gave different people different combinations of clothes, different bags, and different handheld items. We asked each person to take about a dozen sets of photos in four different photo poses. In total, we took about 1200 photos. We had hoped that by having different clothes and objects, the models would think we had more people. However, the experimental results showed that the results could easily be overfitted if there were fewer people. Meanwhile, video recordings in public places can involve privacy issues. Existing public datasets have sufficiently large amounts of data with privacy permissions. Therefore, we propose to annotate the images of the CUHK-PEDES dataset into Chinese language to create a Chinese dataset (CUHK-PEDES-CHINESE).

In the PIR model optimization component, we analyzed the approaches used to optimize the model in different existing studies. Firstly, we analyzed the text and image neural network models used in different PIR studies. Cao *et al.* (2020) trained a CNN image model for PIR in their study. Liang *et al.* (2022) trained a BERT model for PIR in their study. Chen *et al.* (2021) proposed a cross-modal PIR model (TIPCB) by combining a ResNet50 image model and a BERT-Base-Uncased text model. The TIPCB model outperforms existing PIR model trained in a unimodal environment. Second, we analyzed existing studies on improving the performance of PIR by optimizing the loss function. Li *et al.* (2017) proposed the use of CMCE loss to enhance the performance of the model. Zhang and Lu (2018) proposed CMPC loss to enhance the performance of the model. Therefore, we propose to design a new loss function to improve the performance of the CMCPIR model. Our CPIR model architecture is designed to input an image, train it with a combination of text and image models, and output the final prediction across the modal loss function.

The CPIR model architecture design mainly consists of an image modal processing part and a text modal processing part. In the image modal processing, deep neural networks are used to obtain the appearance feature of pedestrians in the image. In the text modal processing part, deep neural networks are used to extract the key semantic information describing the pedestrians in the text. In addition, there is a fusion module that effectively fuses the image features and text features, establishes the correlation and correspondence between them and outputs the final prediction results. The core of our research is to achieve the CPIR model architecture, which is designed for Chinese pedestrian information retrieval for addressing the representation of the fusion performance after fusion of text and image features. The representation of fused features is represented in this research by designing a novel loss function. The loss function we designed can reasonably reflect the cross-modal information fusion performance and achieve the fusion of Chinese language.

In summary, this study proposes a novel CMCPIR model based on Chinese natural language description (RoberResX4D). In addition, we designed a novel CMCPIR loss function to enhance the performance of our model.

2. Approach

We will introduce the overall architecture of the CMCPIR model network proposed in this study. Firstly, we introduce the process of modifying the PIR English dataset into the CPIR Chinese dataset. Then, we introduce the two important components for our CMCPIR model: the image and the text neural network models.

In image-based pedestrian information retrieval, existing methods do not use Chinese for pedestrian information retrieval. The limitation of the existing research lies in the absence of evaluation parameters adapted to Chinese language and Chinese text-image feature fusion performance. Therefore, a group of highly performing image processing and Sinitic languages-based text processing deep learning models are selected based on extensive literature

review. At the same time, we further endeavor on trying to construct our cross-modal RoberResX4D model for CPIR by modifying and changing the combination of these image processing and text processing models. Our CMCPIR model will be compared to a number of different CMCPIR designs, including a baseline constructed by using the design from a top performing English PIR network model. Finally, a novel loss function is proposed for further performance optimization. A diagram of our research method and experiments is shown in Figure 1.

As shown in Figure 1, the training of the CMCPIR model is divided into three main parts: the construction of the Chinese dataset, the modification of the image and text models for the CMCPIR model, and the design of the loss function. CUHK-PEDES-CHINESE is our modified Chinese PIR dataset. In the selection of text and image models, we propose to validate the ResNet50, ResX4D and EB4 deep image neural network models and the HCMacB, BBC and HCRoBerWE deep text neural network models in the CMCPIR task. In designing the loss function for CMCPIR models, we incorporate the properties of MSE (Zhang & Lu, 2018) and propose a novel mean-square weight cross-modal projection matching error (MSWCMPME) loss function.

2.1 Design of CMCPIR model architecture

In this study, a novel CMCPIR model is built, and the overall details of our CMCPIR model network are shown in Figure 2. This novel CMCPIR model differs from the baseline model in the use of a different text model, a different image model and a different loss function.

As shown in Figure 2, the details of our CMCPIR model in this study consist of an image neural network model and a text neural network model, which are HCRoBerWE model and ResX4D model. In addition, we proposed the MSWCMPME loss function to improve the performance and accuracy of our CMCPIR model.

2.2 Chinese text pre-training model candidates

We analyzed different neural networks used for text feature extraction in existing studies. In the study of Islam (2020), it was found that most of the existing research on Natural Language Processing (NLP) uses the BERT model for text feature extraction. Li, Pei, Li, Luo, and Peng (2020) used BBC to extract Chinese text features in their study. Yufeng and Guoxiu (2023) in their study proposed the HCRoBerWE model to enhance the performance of text feature extraction. These Chinese text models are worth training in the CMCPIR task (Qu, Nguyen, & Shimazu, 2016). Therefore, we proposed to adopt and modify BBC, HCRoBerWE and HCMacB, which perform better in existing English PIR studies, for text feature extraction model candidates in our CMCPIR model.

2.3 Image pre-training model candidates

We analyzed different neural networks used for image feature extraction in existing PIR studies. Ding *et al.* (2021) used the VGG16 model as the main neural network for visual pedestrian information extraction. Zhu *et al*. (2021) and Chen *et al.* (2022) achieved the extraction and matching of the pedestrian information using ResNet50. Moghaddam, Charmi, and Hassanpoor (2023) achieved extracting the attributes of pedestrians using the EfficientNets B4 model. R, Bharamagoudra, Reddy, and Sravani (2023) showed that ResX4D model has better feature extraction accuracy than ResNet. In addition, the lightweight and efficient EfficientNets model is becoming the choice of more researchers. Therefore, we proposed to adopt and modify ResNet50, EB4 and ResX4D, which perform better in existing studies, for image feature extraction model candidates in our CMCPIR model.

2.4 Loss function

The CMCPIR model is based on a dual network to measure the similarity of different pedestrian images and the similarity of different text descriptions with pedestrian images. Therefore, the feature information to be extracted by the CMCPIR model contains the common features of two different modal types of information. We propose MSWCMPME loss to improve the performance of CMCPIR model in this study. The goal of the MSWCMPME loss function is to improve the performance of the CMCPIR task by learning the degree of correspondence between images and text.

The MSWCMPME loss function is a modification of the CMPM loss function that combines the advantages of MSE. Our CMCPIR model is trained by a combination of text and image models. In this study, we mark the text input samples as T and the image input samples as I, with the corresponding labels as L. Firstly, we record the predictions PT and PI resulting from the training of the text neural network model and the image neural network model. Meanwhile, when inputs T and I go through their respective models, we calculated the actual text features (TF) and image features (IF). The next step is to calculate the normalized representations of TF_{norm}, IF_{norm}, and, L_{norm} as shown in Equations (1-3).

$$TF_{norm} = \frac{T}{|T| \times L1} \tag{1}$$

$$IF_{norm} = \frac{I}{|I| \times L1} \tag{2}$$

$$L_{norm} = \frac{L}{|L| \times L1} \tag{3}$$

In these equations, L1 stands for normalization using the L1 paradigm. The L1 paradigm takes the sum of absolute values of elements in a vector. The L2 paradigm takes the sum of squares of elements of a vector and then the square root. The paradigm ratio (L1/L2) is probably around 67, with small variations depending on the pixel. Compared with the original L2 paradigm, the L1 paradigm has better robustness and is more compatible with our CMCPIR task. Then, we calculate the MSWCMPME loss for individual image projections to text (L_{I2T}), the normalized prediction result for image projections to text (P_{I2T}), the MSWCMPME loss for text projections to image (L_{T2I}), and the normalized prediction result for text projections to image (P_{T2I}), as shown in Equations (4-7).

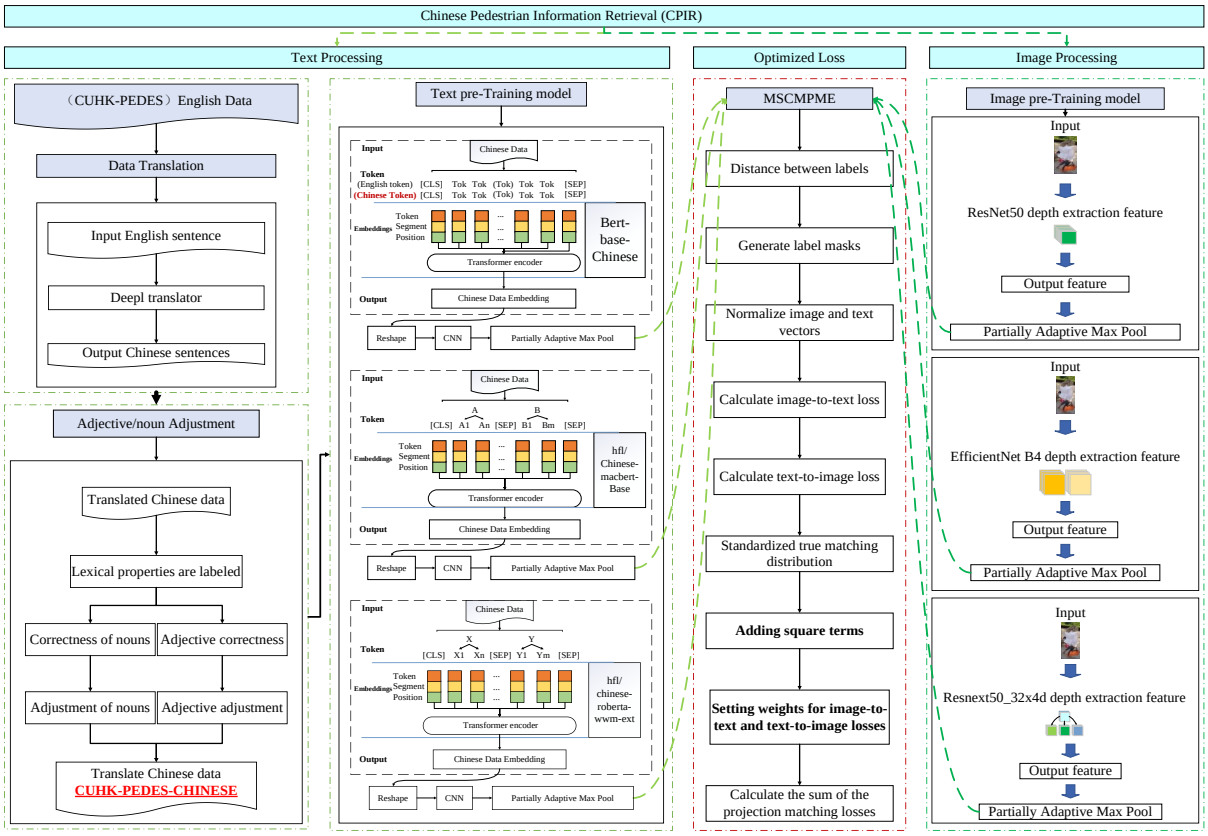

Figure 1. The flowchart of our Chinese pedestrian information retrieval model

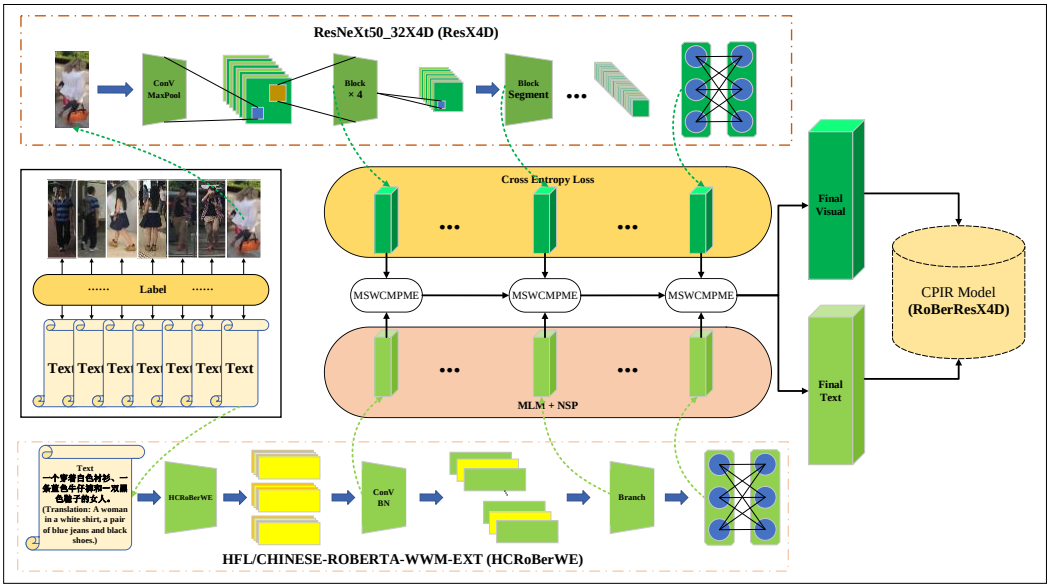

Figure 2. The details of our Chinese Pedestrian Information Retrieval Model

$$P_{I2T} = \frac{T}{|PT| \times L1} \quad (4)$$

$$L_{I2T} = \alpha \times (P_{I2T} \times (P_{I2T} - log(TF_{norm} + e)) + \beta \times (P_{I2T} \times (P_{I2T} - log(TF_{norm} + e))^2 \quad (5)$$

$$P_{T2I} = \frac{I}{|PI| \times L1} \quad (6)$$

$$L_{T2I} = \alpha \times (P_{T2I} \times (P_{T2I} - log(IF_{norm} + e)) + \beta \times (P_{T2I} \times (P_{T2I} - log(IF_{norm} + e))^2 \quad (7)$$

Here α and β are weighting parameters, $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$. We added the squared loss term to the L_{T2I} and L_{I2T} of the internal losses. The e refers to the base of natural logarithms, which has a value of about 2.71828. The reason for adding e is to prevent the logarithmic function from becoming negative, making it impossible for the gradient to descend. Firstly, we increased the penalization of the model by adding $(P_{T2I} \text{ x } (P_{T2I} - log(IF_{norm} + e))^2$ and $(P_{I2T} \text{ x } (P_{I2T} - log(TF_{norm} + e))^2$. We want the model to be more confident in making of predictions. The squared loss term makes the model more sensitive to the size of the error. Small changes in error can lead to large increases in losses. Adding the squared loss term strengthens the matching loss term and the model is more accurate in making predictions. At the same time, adding the squared loss term mitigates the effect of outlier samples on the loss to some extent and prevents gradient explosion. Secondly, we use two weighting parameters α and β. We add the weighting parameters to make it easier to adjust the contribution of different loss terms to the CMCPIR model. With different weight ratios, we can control the degree of influence of the matching loss and the square loss term. By adjusting the weight parameters, we can optimize the matching performance of the model for different application scenarios and datasets. Our experimental results show that a value of 0.5 for α and 0.1 for β are the most suitable for our CMCPIR model. We use log loss to predict the difference between the probability distribution and the label distribution of each normalized sample data. Adding the log-loss computational item can help make the predictions of the model closer to the true label distribution, thus improving the matching performance. In addition, N is the number of image-text pairs. Finally, we summed all the L_{I2T} and L_{T2I} separately, and then averaged the L_{I2T} and L_{T2I} via sums to obtain the final Loss of MSWCMPME (L_{MSWCMPME}):

$$L_{\text{MSWCMPME}} = \frac{1}{N}\sum_{i=1}^{N} L_{I2T} + \frac{1}{N}\sum_{i=1}^{N} L_{T2I} \qquad (8)$$

In order to visualize the nature of the loss function, we try to use geometric images to represent the existing loss function and the MSWCMPME loss function. The loss function images are shown in Figure 3.

As shown in Figure 3, the MSWCMPME loss function is a space-varying function. The MSWCMPME has optimal gradient in the neighborhood of the Z-axis, which makes it easier for the MSWCMPME loss function to converge. Meanwhile, the MSWCMPME loss function computes the gradient descent from space to a point and converges faster than the CMPM loss function. The MSWCMPME loss function is more dimensional and more efficient than the CMPM loss function.

3. Results and Discussion

In this section, we experimentally validate the CMCPIR model performance in terms of CMCPIR model, loss function and actual retrieval performance.

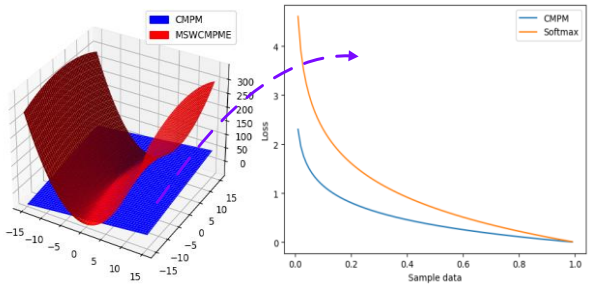

Figure 3. Comparison of our loss function against existing loss functions

3.1 Experimental setup

We divided the experimental setup into the construction of the dataset, the model evaluation metrics, the basic hyperparameter settings, and the analysis of the experimental results.

3.1.1 Dataset

We translated the English descriptions of the CUHK-PEDES dataset into Chinese by machine translation. We hired two language experts to check the translated sentences. The experts focused on the placement of adjectives and the correctness of nouns. We checked a total of 80,416 sentences, and the entire annotation process took about 5 months. Then, we adjusted the Chinese encoding to match our model. We call the Chinese-labeled dataset CUHK-PEDES-CHINESE.

3.1.2 Evaluation metric

The commonly used quantitative evaluation metrics in existing studies in the field of PIR are rank-k accuracy and mean accuracy (mAP) (Suo *et al.*, 2022). In the model testing phase, we categorize the retrieved information into six key descriptors for scoring (see Section 3.5 for details). Finally, we calculate the accuracy of CPIR based on the accuracy of the key descriptors.

3.1.3 Experimental environment and baseline model

In this research, a new CMCPIR model is proposed by comparing and modifying the combination of three text models and three image models. We derive the most suitable hyperparameters based on the limitations of computer computing power, as shown in Table 1. We trained a baseline HCMacBRes50 model using the HCMacB Chinese text model, the ResNet50 image model and the CMPM loss function in the above training environment.

3.2 Results of text and image model modification using baseline model

We replaced the text and image models of the HCMacBRes50 model using the control variable method. The training results of the models are shown in Figure 4.

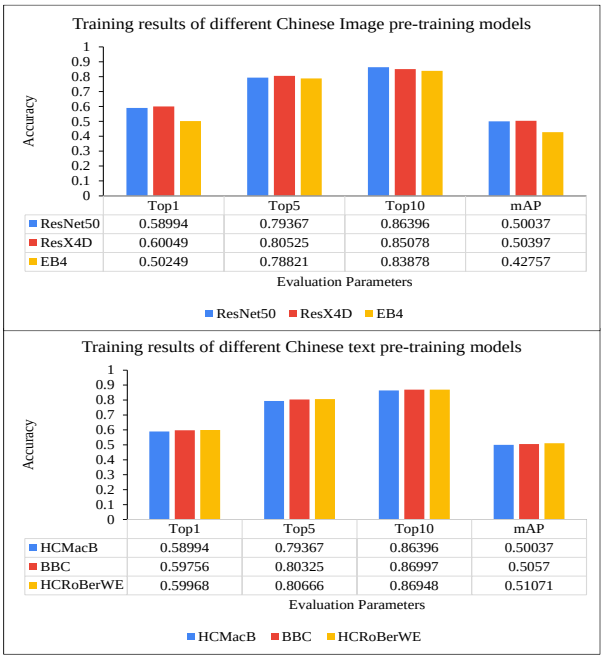

Figure 4. Results of modifying text and image models using baseline model

3.3 Results of different designs of CMCPIR models

As shown in Table 2, the cross-combination of text and image models greatly improves the performance of the CMCPIR model.

3.4 Comparison of loss functions

Table 3 shows the results of training different designs of CMCPIR models using CMPM and MSWCMPME loss functions. The MSWCMPME loss function enhances the CMCPIR model.

3.5 Analysis of actual retrieval results of CMCPIR models

As shown in Figure 5, English language was used as the original HKU dataset labeling text. We use Chinese translations of the English labelling of the original dataset.

The retrieval results of the RoberResX4D model are shown in Table 4. As shown in Table 4, we recorded whether the six key description points matched the CPIR tasks and counted the final scores of the CPIR tasks.

Table 5 shows the key descriptor scores for the validated baseline, RoberResX4D and BBCRes50 models.

3.6 Comparison of retrieval results of existing pedestrian information retrieval models

Most existing research does not support models for pedestrian information retrieval using Chinese. Therefore, we directly compare the retrieval performance without distinguishing the language. We selected SRCF, SAF and TIPCB English pedestrian information retrieval models based on the retrieval ranking of the CUHK dataset. The rankings of the pedestrian information retrieval models are available at https://paperswithcode.com/sota/nlp-based-person-retrival-on-cuhk-pedes?tag_filter=10. Also, we used the same computer

Table 1. Experimental setup and results of a baseline Chinese pedestrian information retrieval model

Pedestrian information retrieval model overview		Experimental environment		Hyper-parameter settings	
Text models	HCMacB; BBC; HCRoBerWE	Software and hardware	Version	Epoch	80
Image models	ResNet50; ResX4D; EB4	PyTorch	1.9.4	Batch Size	64
Loss function	CMPM; MSWCMPME	Python	3.6	Optimizer	ADAM
Dataset	CUHK-PEDES-CHINESE	Colab Pro +	GPU A100	Weights	0.1
Input	Image(384 × 128 × 3) + Text	CUDA	12	Regularization parameter	Lasso
Output	Mapping of pedestrians and descriptions	Transformers	4.30.2	Learning rate	0.003

Baseline pedestrian information retrieval model training results

Text model	Image model	Loss function	Top1	Top5	Top10	mAP
HCMacB	ResNet50	CMPM	0.58994	0.79367	0.86396	0.50037

Table 2. Cross combination of image model and text model for CPIR

Text model	Image model	Top1	Top5	Top10	mAP
HCMacB	ResNet50	0.58994	0.79367	0.86396	0.50037
	ResX4D	0.59949	0.80525	0.85078	0.50397
	EB4	0.50249	0.78821	0.83878	0.42757
BBC	ResNet50	0.59756	0.80325	0.86997	0.5057
	ResX4D	0.60049	0.80195	0.87078	0.50389
	EB4	0.51494	0.74221	0.82078	0.43377
HCRoBerWE	ResNet50	0.59968	0.80666	0.86948	0.51071
	ResX4D	0.60109	0.80295	0.87082	0.50934
	EB4	0.50382	0.73217	0.82078	0.44757

Table 3. Comparison of loss functions for CPIR

Text model	Image model	Loss	Top1	Top5	Top10	mAP
HCMacB	ResNet50	CMPM	0.58994	0.79367	0.86396	0.50037
		MSWTPME	0.59303	0.80675	0.86862	0.50567
	ResX4D	CMPM	0.58995	0.80765	0.86965	0.50678
		MSWTPME	0.59502	0.80998	0.86995	0.50876
	EB4	CMPM	0.52889	0.72653	0.82746	0.44564
		MSWTPME	0.54809	0.75986	0.84568	0.45672
BBC	ResNet50	CMPM	0.59756	0.80025	0.86997	0.5057
		MSWTPME	0.59954	0.80567	0.86789	0.50789
	ResX4D	CMPM	0.60049	0.80195	0.87078	0.50389
		MSWTPME	0.60332	0.80976	0.87897	0.50986
	EB4	CMPM	0.51494	0.74221	0.82078	0.43377
		MSWTPME	0.55903	0.78965	0.84562	0.45621
HCRoBerWE	ResNet50	CMPM	0.59968	0.80666	0.86948	0.51071
		MSWTPME	0.59989	0.80934	0.86908	0.51295
	ResX4D	CMPM	0.60278	0.80995	0.86978	0.51099
		MSWTPME	0.61623	0.81023	0.87516	0.51364
	EB4	CMPM	0.55875	0.78697	0.82234	0.45342
		MSWTPME	0.57985	0.80965	0.85632	0.48756

Table 4. The actual pedestrian information retrieval results of the RoberResX4D model

RoBerResX4D (Ours)							
Search task	Upper part of the body			Lower part of the body			Score
	Hair (long or short)	Clothes (long or short)	Clothes (color)	Trouser (long or short)	Trouser (color)	Shoes (color)	
1	√	√	√	√	√	√	6
2	√	×	√	×	√	√	4
3	√	√	×	√	√	√	5
4	√	√	√	√	√	√	6
5	√	√	√	√	√	√	6
6	√	√	√	√	√	√	6
7	√	√	√	√	√	√	6
8	×	√	√	√	√	√	5
9	×	√	√	√	√	×	4
10	√	√	√	√	√	√	6
11	√	√	√	√	√	√	6
12	√	×	√	×	√	×	3
13	√	√	√	√	√	√	6
14	√	√	√	√	√	√	5
15	√	√	√	√	√	√	6
16	×	√	√	√	√	√	5
17	√	√	√	√	√	×	5
18	√	√	√	√	√	√	6
19	√	√	×	√	√	√	5
20	√	√	√	√	√	√	6
21	√	√	√	√	√	√	6
22	√	×	√	×	√	√	4
23	√	√	×	√	√	√	5
24	√	√	√	√	√	√	6
25	√	√	√	√	√	√	6
26	√	√	√	√	√	√	6
27	√	√	√	√	√	√	6
28	×	√	√	√	√	×	5
29	√	√	√	√	√	√	6
30	√	√	√	√	√	√	6
31	√	√	√	√	√	√	6
32	√	×	√	×	√	×	3
33	√	√	√	√	√	√	6
34	√	√	√	√	√	√	6

Table 4. Continued.

RoBerResX4D (Ours)							
Search task	Upper part of the body			Lower part of the body			Score
	Hair (long or short)	Clothes (long or short)	Clothes (color)	Trouser (long or short)	Trouser (color)	Shoes (color)	
35	√	√	√	√	√	√	6
36	×	√	√	√	√	√	5
37	√	√	√	√	√	×	5
38	√	√	√	√	√	√	6
39	√	√	√	√	√	√	6
40	√	√	√	√	√	√	6
41	√	√	√	√	√	√	6
42	√	√	√	√	√	√	6
43	√	√	×	√	√	√	5
44	√	√	√	√	√	√	6
45	√	√	√	√	√	√	6
46	×	√	×	√	×	√	3
47	√	√	√	√	√	√	6
48	×	√	√	√	√	√	6
49	×	√	√	√	√	×	4
50	√	√	√	√	√	√	6

Table 5. The actual retrieval performance and accuracy rate for key descriptors of different pedestrian information retrieval models

Model		HCMacBRes50 (Baseline)	BBCRes50 (Ours)	RoBerResX4D (Ours)
Score	1	6%	2%	0%
	2	8%	8%	0%
	3	16%	4%	6%
	4	4%	16%	8%
	5	10%	14%	22%
	6	54%	56%	64%
Key descriptors	Hair (long or short)	72%	78%	80%
	Clothes (long or short)	74%	86%	88%
	Clothes (color)	88%	86%	88%
	trouser (long or short)	86%	82%	86%
	trouser (color)	90%	88%	88%
	Shoes (color)	70%	80%	84%

Figure 5. Examples of visualization of pedestrian information retrieval model for actual retrieval results

Table 6. Actual retrieval performance and accuracy of key descriptors of existing english pedestrian information retrieval models

Model		SRCF(Top1)	SAF(Top2)	TIPCB(Top3)
Score	1	0%	0%	0%
	2	0%	0%	0%
	3	6%	14%	6%
	400%	10%	8%	14%
	500%	20%	16%	18%
	600%	64%	62%	62%
Key Descriptors	Hair (long or short)	82%	84%	78%
	Clothes (long or short)	84%	82%	76%
	Clothes (color)	90%	88%	84%
	trouser (long or short)	82%	80%	78%
	trouser (color)	88%	84%	86%
	Shoes (color)	90%	88%	84%

and parameter settings for training. The performances of different models for pedestrian information retrieval are shown in Table 6.

3.7 Analysis of experimental results

Throughout Table 5 and Table 6, our proposed model RoBerResX4D achieves the same retrieval performance as SRCF. Meanwhile, Table 3 shows that the use of MSWCMPME loss function leads to an improved key descriptor accuracy. MSWCMPME loss function can effectively improve the robustness of the model.

4. Conclusions

We investigated a CMCPIR model based on Chinese natural language description. Firstly, we use the original English CUHK-PEDES dataset to construct the Chinese-labelled CUHK-PEDES-CHINESE dataset to solve the absence of data problem for the Chinese CPIR task. Second, we screened different deep learning models for image processing and text processing. We constructed novel cross-modal CPIR models by combining different image and text models. Thirdly, we propose a new MSWCMPME loss function to further optimize the performance of the cross-modal RoberResX4D model.

In this research, we propose a novel RoberResX4D Chinese language pedestrian information retrieval model. RoberResX4D can use Chinese language for pedestrian information retrieval. In addition, we propose a novel MSWCMPME loss function to improve the performance of the model, and the pedestrian information retrieval performance of RoberResX4D model can reach the performance of the existing SRCF model. Meanwhile, the proposed MSWCMPME loss function improves the robustness of RoberResX4D model.

Meanwhile, by comparing the performance with existing English pedestrian information retrieval models, we analyze that the problem may arise in the translation of the dataset. Our practice of retaining only a portion of the adjectives resulted in a decrease in the accuracy of the key descriptors. In the future, the model can be trained using the Chinese dataset with more adjectives to expect an improvement in the performance of its descriptive information retrieval model.

Code and experiments of this work have been uploaded to Github https://github.com/Axieyan/CPIR-RoberResX4D.

Acknowledgements

Conceptualization, Y.X. and J. Q.; methodology, Y.X. and J. Q.; software, Y.X. and J. Q.; validation, Y.X. and J. Q.; formal analysis, Y.X. and J. Q.; investigation, Y.X. and J. Q.; data curation, Y.X. and J. Q.; writing—original draft preparation, Y.X. and J. Q.; writing—review and editing, Y.X. and J. Q.; visualization, Y.X. and J. Q.; supervision, J. Q.; All authors have read and agreed to the published version of the manuscript.

References

Cao, B., Araújo, A. F., & Sim, J. (2020). Unifying deep local and global features for image search. *European Conference on Computer Vision*. Retrieved from https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123650715.pdf

Chandrakala, M. V., & Devi, P. D. (2021). Two-stage classifier for face recognition using hog features. *Materials Today: Proceedings*. Retrieved from https://doi.org/10.1016/j.matpr.2021.04.114

Chen, Y., Zhang, G., Lu, Y., Wang, Z., Zheng, Y., & Wang, R. (2021). Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494, 171-181. Retrieved from https://arxiv.org/pdf/2105.11628v1.pdf

Ding, Z., Ding, C., Shao, Z., & Tao, D. (2021). Semantically self-aligned network for text-to-image part-aware person re-identification. *ArXiv*. Retrieved from https://arxiv.org/pdf/2107.12666

Farooq, A., Awais, M., Kittler, J., & Khalid, S. S. (2021). Axm-net: Implicit cross-modal feature alignment for person re-identification. *AAAI Conference on Artificial Intelligence*. Retrieved from https://cdn.aaai.org/ojs/20370/20370-13-24383-1-2-20220628.pdf

Islam, K. (2020). Person search: new paradigm of person re-identification: A survey and outlook of recent works. *Image and Vision Computing, 101,* 103970. Retrieved from https://doi.org/10.1016/j.imavis.2020.103970

Li, C., Pei, Z., Li, L., Luo, Z., & Peng, D. (2020). Research on chinese text semantic matching base on bert and bilstm_attention. *DEStech Transactions on Engineering and Technology Research*. Retrieved from https://doi.org/10.12783/DTETR/MCAEE2020/35027

Li, Q., & Qu, J. (2022). A novel bnb-no-bk method for detecting fraudulent crowdfunding projects. *Songklanakarin Journal of Science and Technology*, *44*(5). Retrieved from https://sjst.psu.ac.th/journal/44-5/7.pdf

Li, S., Xiao, T., Li, H., Yang, W., & Wang, X. (2017). Identity-aware textual-visual matching with latent co-attention. *2017 IEEE International Conference on Computer Vision (ICCV)*, 1908-1917. Retrieved from https://openaccess.thecvf.com/content_ICCV_2017/papers/Li_Identity-Aware_Textual-Visual_Matching_ICCV_2017_paper.pdf

Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., & Wang, X. (2017). Person search with natural language description. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5187-5196. Retrieved from https://openaccess.thecvf.com/content_cvpr_2017/papers/Li_Person_Search_With_CVPR_2017_paper.pdf

Liang, T., Lin, G., Wan, M., Li, T., Ma, G., & Lv, F. (2022). Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15471-15480. Retrieved from https://openaccess.thecvf.com/content/CVPR2022/papers/Liang_Expanding_Large_Pre-Trained_Unimodal_Models_With_Multimodal_Information_Injection_for_CVPR_2022_paper.pdf

Moghaddam, M., Charmi, M., & Hassanpoor, H. (2023). A robust attribute-aware and real-time multi-target multi-camera tracking system using multi-scale enriched features and hierarchical clustering. *Journal of Real-Time Image Processing*, *20*, 1-14. Retrieved from https://doi.org/10.1007/s11554-023-01301-y

Qu, J., Nguyen, L., & Shimazu, A. (2016). Cross-language information extraction and auto evaluation for oov term translations. *China Communications*, *13*, 277-296. Retrieved from https://doi.org/10.1109/CC.2016.7897550

R, L., Bharamagoudra, M. R., Reddy, T. S., & Sravani, K. (2023). Performance analysis of convolutional neural network for plant diseases identification. *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, 1-7. Retrieved from https://doi.org/10.1109/I2CT57861.2023.10126398

Suo, W., Sun, M., Niu, K., Gao, Y., Wang, P., Zhang, Y., & Wu, Q. (2022). A simple and robust correlation filtering method for text-based person search. *European Conference on Computer Vision*. Retrieved from https://doi.org/10.1007/978-3-031-19833-5_42

Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., Ding, E., & Huang, J. (2021). Dolg: single-stage image retrieval with deep orthogonal fusion of local and global features. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11752-11761. Retrieved from https://doi.org/10.1109/ICCV48922.2021.01156

Yufeng, D., & Guoxiu, H. (2023). Analysis of neural network modules for named entity recognition of chinese medical texts. *Data Analysis and Knowledge Discovery*. Retrieved from https://doi.org/10.11925/infotech.2096-3467.2022.0908

Zhang, Y., & Lu, H. (2018). Deep cross-modal projection learning for image-text matching. *European Conference on Computer Vision*. Retrieved from https://openaccess.thecvf.com/content_ECCV_2018/papers/Ying_Zhang_Deep_Cross-Modal_Projection_ECCV_2018_paper.pdf

Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., & Hua, G. (2021). Dssl: deep surroundings-person separation learning for text-based person retrieval. *Proceedings of the 29th ACM International Conference on Multimedia*. Retrieved from https://arxiv.org/pdf/2109.05534