Hindawi
Mobile Information Systems
Volume 2021, Article ID 5512382, 7 pages
https://doi.org/10.1155/2021/5512382

Research Article

Multilabel CNN-Based Hybrid Learning Metric for Pedestrian Reidentification

Yinjun Zhang,[1] **Ryan Alturki**,[2] **Hasan J. Alyamani**,[3] **Mohammed Abdulaziz Ikram**,[4] **Ateeq ur Rehman**,[5] **and Muhammad Haleem**[6]

[1]*School of Mechanical and Electrical Engineering, Guangxi Science & Technology Normal University, Laibing, China*
[2]*Department of Information Science, College of Computer and Information Systems, Umm Al-Qura University, Mecca, Saudi Arabia*
[3]*Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Saudi Arabia*
[4]*Computer Science Department, University College in Al-Jamoum, Umm Al-Qura University, Mecca, Saudi Arabia*
[5]*Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan*
[6]*Department of Computer Science, Faculty of Engineering and Technology, Kardan University Kabul, Kabul, Afghanistan*

Correspondence should be addressed to Muhammad Haleem; m.haleem@kardan.edu.af

Received 7 February 2021; Revised 8 March 2021; Accepted 24 March 2021; Published 5 April 2021

Academic Editor: Nicola Bicocchi

Pedestrian reidentification has recently emerged as a hot topic that attains considerable attention since it can be applied to many potential applications in the surveillance system. However, high-accuracy pedestrian reidentification is a stimulating research problem because of variations in viewpoints, color, light, and other reasons. This work addresses the interferences and improves pedestrian reidentification accuracy by proposing two novel algorithms, pedestrian multilabel learning, and investigating hybrid learning metrics. First, unlike the existing models, we construct the identification framework using two subnetworks, namely, part detection subnetwork and feature extraction subnetwork, to obtain pedestrian attributes and low-level feature scores, respectively. Then, a hybrid learning metric that combines pedestrian attributes and low-level feature scores is proposed. Both low-level features and pedestrian attributes are utilized, thus enhancing the identification rate. Our simulation results on both datasets, i.e., CUHK03 and VIPeR, reveal that the identification rate is improved compared to the existing pedestrian reidentification methods.

1. Introduction

In recent years, pedestrian reidentification emerges as a hot research topic. It has attained considerable attention since it can be applied to many potential applications in human-computer interaction and surveillance tasks. The purpose of pedestrian reidentification algorithms is to search and detect a target from a large set of images. Various cameras capture these images for detecting a target image, where an image or a video sequence can represent the target person. The pedestrian reidentification algorithm is imperative in the development of an automatic video surveillance system. Pedestrian reidentification is an especially difficult topic due to tremendous variations in viewpoints, human poses, light, and other factors. These impacting factors are likely to make two independent images of a single person look quite different. On the contrary, these factors can also make some people's images look pretty similar, creating great difficulties for the identification algorithms. The previous works [1, 2] mainly aim to eliminate the impacts of interference and build a reliable pedestrian reidentification algorithm with solid robustness.

The existing pedestrian reidentification algorithms have two major types, namely, supervised learning and unsupervised learning. For supervised learning-based approaches [3, 4], the work presented in [1] used a deep convolutional neural network (CNN) for learning features continuously, and the resultant matching metrics used in reidentification of individuals expressively increase the precision of the state-of-the-art model. The authors in [3] presented indiscriminative reinforcement lean strategies and multi-instance multilabel learning methods to solve the pedestrian reidentification within a short-term surveillance system. As for unsupervised learning-based pedestrian reidentification algorithms [5–7], the authors in [5] presented a progressive unsupervised learning (PUL) method for transferring the pretrained model's deep representations to unseen domains. In [6], a transferable approach simultaneously learnt the attribute-semantic and identity-discriminative feature space transferred to new fields for reidentification tasks without new labeled data.

The first significant challenge in designing pedestrian reidentification algorithms is the weak capability to effectively extract and match the features from different views and select mutual patterns. However, one image's features may not necessarily appear the same in the target image due to the pose change caused by view difference [8, 9]. Hence, extracting, embedding, and evaluating features from different image domains are critical in developing a high-performance pedestrian reidentification framework. The other challenge is how to select a good metric that measures the similarity of various features within sampled images, which is the most crucial part of the training process of CNN models. A good metric can increase the learning efficiency of CNN models during training, thus improving the recognition performance.

This work improves pedestrian reidentification by adopting pedestrian multilabel learning and investigating hybrid learning metrics. The contributions of this paper are twofold. First, two subnetworks, namely, part detection subnetwork and feature extraction subnetwork, are used to obtain pedestrian attributes and low-level feature scores. A hybrid learning metric that combines pedestrian attributes and low-level feature scores is proposed to enhance the performance. The experiment results validate the performance of the proposed framework.

The remainder of this paper is in accordance with the following pattern. Section 2 introduces previous works related to pedestrian reidentification. Section 3 describes the proposed multilabel learning algorithms with enhanced learning metrics. Section 4 illustrates the experimental setup and experimental results to validate the performance of the proposed algorithms. Finally, Section 5 concludes the paper.

2. Related Work

This section describes the literature from two different perspectives: pedestrian reidentification using traditional approaches and pedestrian identification using machine learning approaches.

2.1. Pedestrian Reidentification. Pedestrian reidentification can be viewed as the problem of identifying pedestrians from different images captured by different cameras. Figure 1 illustrates some example images of pedestrians captured by using real cameras in the CUHK03 [9] dataset. Pedestrian reidentification typically involves two parts: extracting features from input images and comparing the extracted features' metrics to obtain the rank list. From these two aspects, the previous works on pedestrian reidentification mainly aim at two factors: developing new learning metrics [10, 11] and developing new feature representations [12]. These schemes are used to combat the variations of viewpoint, pose, and color.

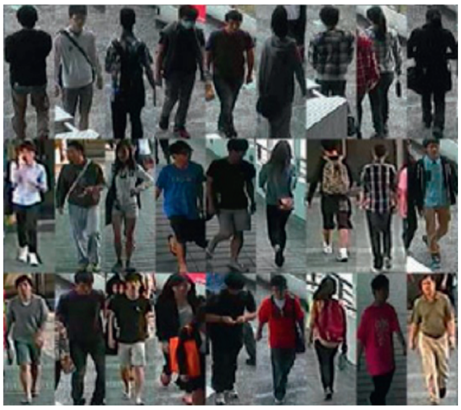

Figure 1: Example images of pedestrians from the CUHK03 [9] dataset.

2.2. Deep Learning-Based Pedestrian Reidentification. With deep learning algorithms' resurgence, deep neural networks [7, 13] have witnessed great success in many fields, especially computer vision tasks. CNNs can extract low-level features while learning more abstract information, including detailed texture and geometry patterns. Based on self-built CNN models, several deep learning methods [1, 2, 9] have been presented to realize pedestrian reidentification tasks and have shown promising performance improvement compared to traditional handcrafted features.

3. The Proposed Pedestrian Multilabel Learning Algorithms

This section introduces proposed algorithms for pedestrian multilabel learning, which includes a part detection subnetwork. Then, the other subnetwork performs feature extraction for the input and evaluates the similarity of two given images. As mentioned above, both subnetworks receive the input paired images and process them in a parallel manner. The following sections describe associated algorithms in detail.

3.1. Overview. The problem of pedestrian reidentification can be described in a similar way to object recognition. The conventional reidentification algorithm takes two input images; each image usually involves the pedestrian's whole body. The identification algorithm outputs the similarity metric between the given images, which shows the probability that the two images depict the same or different people.

The existing frameworks [1, 2, 4, 5] typically utilize bottom-up image cues and end-to-end learning for reidentification. These works improve the performance and accuracy of pedestrian reidentification mainly through two aspects. The first aspect is to design learning architectures

[1, 2], thus enhancing the feature extraction abilities of reidentification models. The second aspect is to develop efficient evaluation metrics [4, 5] and precisely evaluate the distance between correct and trained results, thereby improving the learning capabilities.

One critical problem that limits the performance of conventional pedestrian reidentification is that these works fail to utilize the small semantic parts that contain crucial information of pedestrian attributes, such as hair color and gender. The conventional pedestrian reidentification algorithms mainly emphasize bottom-up image indications while neglecting small semantic parts, thus losing physical pedestrian attributes. Moreover, some of these attributes have local characteristics. Due to the variation of viewpoints and other impacting factors, the failure of taking advantage of this information may become the performance bottleneck. According to the previous works on semantic part detection [14], the pedestrian's multiple attributes can be used to describe and improve the overall accuracy.

Inspired by the multilabel CNN model [15] and part detection approaches [14], we design a pedestrian multilabel learning framework to tackle fine-grained pedestrian attribute recognition. Figure 2 shows the proposed framework of pedestrian reidentification with the aid of multilabel learning. The processes of proposed pedestrian multilabel learning are carried out as follows:

(1) First, the input image passes the part detection subnetwork, responsible for dividing the predefined body parts. A fully connected (FC) layer will calculate the attribute scores associated with divided body parts after the split body parts are flattened.

(2) Then, the same input image passes the feature extraction subnetwork, which extracts the low features hidden in the image. The low-level feature similarity will be calculated by an FC layer similar to the attribute scores in Step 1.

(3) Finally, after the attribute scores and low-level feature similarity scores are obtained, the gallery images will be ranked according to the two metrics.

3.2. Part Division. According to Zhang et al. [14], body movements and other factors may cause the popularly used universal feature illustration approaches to agonize from misalignments. Hence, combating and eliminating the interference that arose from camera viewpoints' variations becomes a challenging task for body part division tasks. The previous work [15] solved this problem by dividing the given image into 15 overlapping areas. Several softmax classifiers were then used to compute various body parts' regression and then obtain the attribute scores. However, there are two defects in this method. Firstly, the positions of the human body part vary in various images. The handcrafted division of areas may not accommodate all the cases, thus resulting in inaccurate body parts. Secondly, the 15 overlapping body parts need to pass multiple CNNs, which requires a considerable amount of computational complexity and makes this scheme inefficient.

We use another body part division scheme to improve the efficiency as well as accuracy. Similar to [14], we adopt a body part detector, named the part detection subnetwork in Figure 2. The part detection subnetwork integrates the state-of-the-art semantic model, R-CNN [16], to detect each body part region. Detecting every part of the human body is difficult since the resolution of images in an existing dataset, such as CUHK03 [9], is low. Hence, we let the part detection subnetwork only find those parts associated with human hair and clothing. Assuming that there are m parts to be detected, the part detection subnetwork has $(m+1)$ parallel output labels, consisting of m body parts and one gender label, indicating male or female. Based on the loss of R-CNN [16], the part detection subnetwork is trained and optimized using the following multitask loss L_{part}:

$$\mathscr{L}_{\text{part}} = \mathscr{L}_{\text{class}}(s,c) + \lambda[c>0]\cdot\mathscr{L}_{\text{loc}}\left(b^{r}, b^{\text{true}}\right), \tag{1}$$

where $s\in[0,1]$ denotes the confidence score for each regressed bounding box b, which is the output of the part detection subnetwork. $c\in[0,m]$ is the ground-truth class of the body part bounding box, while $\mathscr{L}_{\text{class}}$ represents the loss of the ground-truth class. Besides, $\mathscr{L}_{\text{loc}}$ denotes the loss function for the regressed bounding box for each body part. b^{true} is the ground-truth result, while b^{r} represents the regressed bounding box of the true class.

The FC layer in Figure 2 acts as the classifier for the feature vector of the body part. As the concatenation fully connected layer in [14], we use the following matrix-vector multiplication and nonlinear activation to compute the combination of various body parts and realize fine-grained classification:

$$y = f\left(\sum_{i=1}^{m+1}\mathbf{W}_i\cdot\mathbf{x}_i\right), \tag{2}$$

where $\mathbf{W}_i$ represents the weight matrix for the i^{th} body part. The feature vector of the i^{th} body part is stored in a vector $\mathbf{x}_i$.

3.3. Learning Metrics. Unlike the conventional frameworks that only compare the extracted low-level features, the extracted attribute scores and low-level feature similarity scores from two subnetworks are aggregated to evaluate the given images' similarity. Pedestrian reidentification algorithms will return a list where the gallery images are ranked by their distances between the probe image and the gallery images. The higher the ground-truth gallery images are ranked, the higher the accuracy is achieved.

The total loss function of the proposed framework is composed of two parts. The first one is the softmax function that computes the loss of all pedestrian attributes. The second part is the cost from the low-level feature similarity. The following equation gives the total loss function:

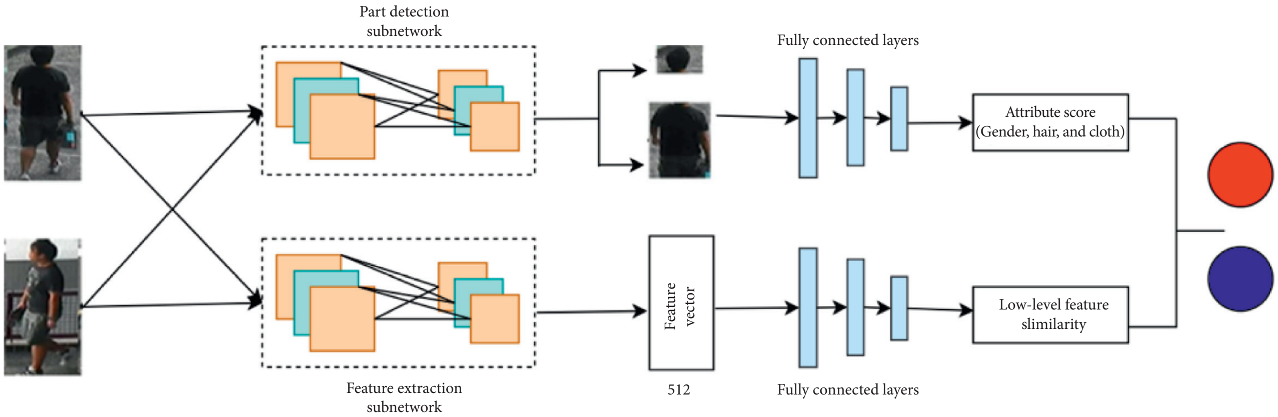
Part detection subnetwork
Fully connected layers
Attribute score (Gender, hair, and cloth)
Feature vector
512
Low-level feature slimilarity
Fully connected layers
Feature extraction subnetwork

FIGURE 2: The proposed framework of pedestrian reidentification with the aid of multilabel learning.

$$\mathscr{L}_{ML} = \sum_{i=1}^{I} \lambda_i L_i + \alpha\left(L_p - L_g\right)^T \left(L_p - L_g\right), \tag{3}$$

where L_i denotes the loss of the i^{th} attribute, while I is the total number of features. λ_i denotes the parameter that defines the contribution of the i^{th} attribute. Without an explicit statement, the value of λ_i is $1/I$ which means each attribute contributes equally. L_p and L_g denote the low-level features of given probe images and gallery images, respectively.

The metric L_i in equation (3) is expanded as follows:

$$L_i = -\frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M^i} 1_{\{y_n^i = m\}} \log \frac{e^{\left(w_m^i\right)^T \cdot x_n^i}}{\sum_{m=1}^{M^i} e^{\left(w_m^i\right)^T \cdot x_n^i}}, \tag{4}$$

where N is the number of total training samples, while M^i denotes the number of classes for the i^{th} attribute. Besides, $\{x_n^i, y_n^i\}$ denotes the paired training image sample, and y_n^i denotes the n^{th} sample x_n^i's i^{th} attribute label.

The proposed multilabel learning algorithms for pedestrian reidentification still reserve the end-to-end learning paradigm as previous works [1, 2, 15]. Therefore, the training and inference phases can be done with high parallelism, thus creating no impact on overall system efficiency. The parallel approach will make the model run smoothly and speed up the proposed model's execution time.

4. Experiments

In this section, we conduct detailed experiments to study the performance of the proposed framework. Besides, the obtained results are compared with other existing pedestrian reidentification approaches to demonstrate our proposed models' superior performance.

4.1. Datasets. We utilize the CUHK03 [9] publicly available dataset to evaluate our work and compare the results with other state-of-the-art models. VIPeR [17] and CUHK03 are the two commonly used person reidentification datasets. VIPeR contains a total of 632 pedestrian images, which are captured by using two cameras with different viewpoints. However, the drawback is that the image resolution of VIPeR is relatively low, making it not suitable for body part division. In contrast, CUHK03 is the pedestrian reidentification dataset large enough to overcome deep neural networks' overfitting. It provides the bounding boxes detected correctly and manually label them. CUHK03 consisted of 13,164 images of various pedestrians and was captured with several surveillance cameras, where every identity is detected with two separate camera views. Such settings yielded an average of 4.8 images in each set of view. Furthermore, the dataset comes up with bounding boxes auto-obtained and manually labeled pedestrian bounding boxes. In our work, we present the results on the labeled dataset. To ensure good detection quality, we select CUHK03 as the test dataset in this paper.

4.2. Implementation Settings. The proposed algorithms and framework are implemented using the famous deep learning library PyTorch [18]. Most functional layers are adopted from the library, and those parts associated with the proposed structure are realized on our own. The training and inference of CNN models are performed on multiple NVIDIA TITAN X GPUs to accelerate the speed. The experiments are conducted by optimizing the proposed softmax-based objective function. We first train the CNN models with the minibatch-based stochastic gradient descent. The minibatch size is set to 16 for smooth gradient updating and convergence. We applied $L2$ regularization and dropout (cite dropout) for the earlier layers with a ratio of 0.5 to avoid overfitting and speed up the convergence speed. Initially, we set the learning rate to 0.05 and used a decreasing factor until we obtained the best results. The models are trained for 24,000 iterations, and the learning rates for two subnetworks are identical (0.001). We use the Adam optimizer [19] to adjust the learning rate and accelerate the optimization convergence.

4.3. Evaluation Protocol. We assume the widely adopted single-shot modality to allow a wide-range comparison with state-of-the-art models. For every probe image, we match it in contradiction to the gallery set to obtain the true match's rank. The rank-k recognition rate describes the match at rank k. At the same time, we record the cumulative values' recognition rate at all ranks due to a specific trail of the cumulative matching characteristic (CMC) result. We evaluated the performance with such settings ten times and reported the average CMC results.

4.4. Model Training. In this section, we discuss the data augmentation, dropout techniques, and training strategies.

4.4.1. Data Augmentation. Even though the selected dataset is large scale, the positive paired data are not as many as opposing pairs. Moreover, the dataset size is relatively small compared to the deep network models. Hence, data imbalance and overfitting may occur. Multiple data augmentation strategies are used to compensate for the performance degradation resulted from insufficient images, overcoming this limitation. Affine transformations are applied to alleviate the overfitting effect. We also augment the dataset by conducting random translations. We sample an equal amount of positive and negative pairs to accomplish data balancing despite creating the negative-positive fixed proportion.

4.4.2. Dropout. In the scenario of person reidentification, because of the considerable misalignment, cross-view variations, occlusions, and pose variations, it is anticipated that specific patches on the identical person (though in various views) may contradict each other. We applied the dropout [20] approach to induce the trained proposed model adequate to misdetection of the similarity. We randomly select some outputs of the first convolutional layer (extracted features with the filter pairs) and set them as zeros at each training iteration and for every training sample as the input. We calculate the gradients in the backpropagation with the randomly muted filter responses to achieve a stable training model.

4.4.3. Training Strategies. It is time-consuming and tedious to train the fast R-CNN model from scratch over large-scale datasets. We use the pretrained fast R-CNN model to avoid complex model tuning and accelerate the training process. The pretrained model significantly reduces the training time and improves the identification accuracy. In the experiment, MobileNetV2 [21] is selected as the feature extraction subnetwork. The MobileNetV2 model uses multiple optimization schemes that achieve a good tradeoff between complexity and performance. The impact of the pretraining effect is also demonstrated experimentally.

4.5. Experiment Results. We compare the result of our model with two person reidentification methods (KISSME [11] and SDALF [22]), two metric learning methods (logistic distance metric learning (LDM) [23] and metric learning to rank (RANK) [24]), and a filter-based approach FPNN [9]. RANK is an optimized approach for ranking-based problems, whereas person reidentification is a ranking problem. LDM is designed for person and face identification scenarios.

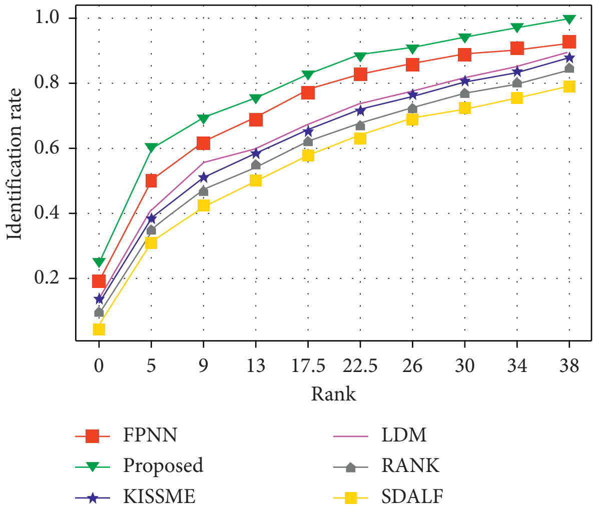

Figure 3: Performance comparison of CMC results with other pedestrian reidentification methods [9, 11] on the CUHK03 dataset.

On the benchmark CUHK03 dataset, we conducted a range of experiments of our proposed and other state-of-the-art models for pedestrian bounding boxes. Figure 3 plots the comparison performance of CMC on various methods on the CUHK03 dataset. It shows that the proposed algorithm significantly outperforms the KISSME [11] model by a large margin with an identification rate improvement of 10% to 18%. As for the other deep neural network-based method, FPNN [9], our approach yields an identification rate improvement by 4% to 8%. The observed performance gain mainly comes from the utilization of pedestrian attribute information.

Since initializing has been instrumental for the performance and convergence of deep learning models [25–27], we propose initializing the proposed model with pretrained weights. Figure 4 presents the performance comparison on the CUHK03 dataset with or without a pretrained subnetwork. From the figure, we can see that the algorithm with a pretrained model achieves a higher identification rate by 2% to 8% for all rank values. The gain comes from the finetuning of the R-CNN model which helps the training start at a good point. Moreover, according to our experiments, the training process with a pretrained model converges faster than that without a pretrained model, significantly reducing the time cost of model training.

The experimental results in Figure 5 demonstrate the usefulness of employing dropout to the proposed model. The figure describes the rank-1 identification rates following various sets of training minibatches on the validation set against a range of dropout rates (0% to 20%). Our model's identification rate decreases as the number of training minibatches grows in the absence of dropout. Such behavior indicates the overfitting problem. The convergence speed improves, and the identification rate is high when the dropout is set to 5%, enabling it to be adequate to patch the misdetection of correspondence and result in a good generalization power. However, it does not achieve a reasonable identification rate when the dropout is set to a higher value (20% in this case).

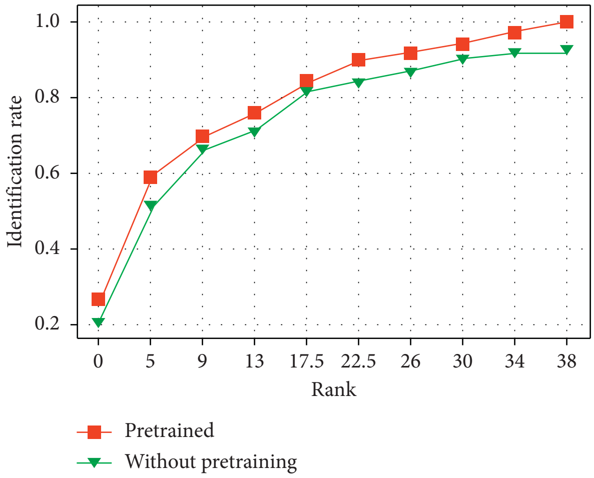

Figure 4: Comparison of CMC results on the CUHK03 dataset with or without the pretrained subnetwork.

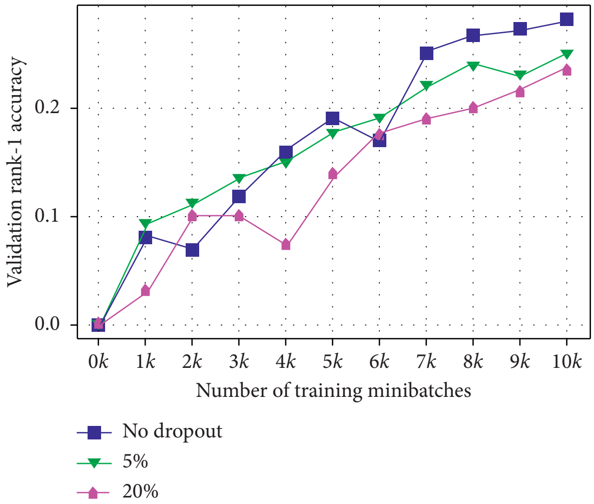

Figure 5: Rank-1 identification of the proposed model on the validation set after a different number of training minibatches.

5. Conclusion

In this paper, we present a novel pedestrian reidentification framework based on multilabel learning CNN models. Besides, we also propose a hybrid algorithm with the aid of pedestrian attributes and low-level features. Both low-level features and pedestrian attributes are utilized to enhance the performance. Experimental results on the popular dataset, CUHK03, show that the identification rate is improved compared to the existing algorithms.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Acknowledgments

This work was sponsored by Chongqing Basic Science and Research Project (CSTC2015JCYJBX0133) and NSFC (51375519).

References

[1] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3908–3916, Boston, MA, USA, June 2015.

[2] R. R. Varior, M. Haloi, and G. Wang, “Gated Siamese convolutional neural network architecture for human re-identification,” in *European Conference on Computer Vision*, pp. 791–808, Springer, Berlin, Germany, 2016.

[3] Y. Lin, F. Guo, L. Cao, and J. Wang, “Person re-identification based on multi-instance multi-label learning,” *Neurocomputing*, vol. 217, pp. 19–26, 2016.

[4] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, “Unsupervised person re-identification by soft multi-label learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2148–2157, Long Beach, CA, USA, 2019.

[5] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person Re-identification,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 4, pp. 1–18, 2018.

[6] J. Wang, X. Zhu, S. Gong, and W. Li, “Transferable joint attribute-identity deep learning for unsupervised person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2275–2284, Salt Lake City, UT, USA, June 2018.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[8] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, pp. 5998–6008, MIT Press, Cambridge, MA, USA, 2017.

[9] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: deep filter pairing neural network for person re-identification,” in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159, Columbus, OH, USA, June 2014.

[10] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197–2206, Boston, MA, USA, June 2015.

[11] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288–2295, Providence, RI, USA, June 2012.

[12] I. Kviatkovsky, A. Adam, and E. Rivlin, “Color invariants for person re-identification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2012.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” 2020, https://arxiv.org/abs/2005.12872.

[14] H. Zhang, T. Xu, M. Elhoseiny et al., “SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1143–1152, Las Vegas, NV, USA, June 2016.

[15] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, “Multi-label CNN based pedestrian attribute learning for soft biometrics,” in *Proceedings of the International Conference on Biometrics (ICB)*, pp. 535–540, Phuket, Thailand, May 2015.

[16] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-CNN: towards real-time object detection with region proposal networks,” in *Advances in Neural Information Processing Systems*, pp. 91–99, MIT Press, Cambridge, MA, USA, 2015.

[17] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *European Conference on Computer Vision*, pp. 262–275, Springer, Berlin, Germany, 2008.

[18] A. Paszke, S. Gross, S. Chintala et al., “Automatic differentiation in pytorch,” in *Proceedings of the Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.

[19] D. P. Kingma and B. Jimmy, “ADAM: a method for stochastic optimization,” 2014, https://arxiv.org/abs/1412.6980.

[20] N. Srivastava, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, June 2018.

[22] M. Farenzena, “Person re-identification by a symmetry-driven accumulation of local features,” in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010.

[23] M. Guillaumin, J. Verbeek, and C. Schmid, “Is that you? Metric learning approaches for face identification,” in *Proceedings of the IEEE 12th International Conference on Computer Vision*.

[24] B. McFee and G. R. G. Lanckriet, “Metric learning to rank,” in *Proceedings of the International Conference on Machine Learning*, Haifa, Israel, June 2010.

[25] D. P. Kingma and J. Ba, “ADAM: a method for stochastic optimization,” 2014, https://arxiv.org/abs/1412.6980.

[26] D. Erhan, “Why does unsupervised pre-training help deep learning?” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, May 2010.

[27] Shah, S. T. Ullah, J. Li, Z. Guo, G. Li, and Q. Zhou, “DDFL: a deep dual function learning-based model for recommender systems,” *Database Systems for Advanced Applications*, Springer, Cham, Switzerland, 2020.