Person Search via Anchor-Free Detection and Part-Based Group Feature Similarity Estimation

Qing Liu, Keyang Cheng(✉), and Bin Wu

School of Computer Science and Communication Engineering,
Jiangsu University, Zhenjiang 212013, China
kycheng@ujs.edu.cn

Abstract. In order to solve the problems of insufficient accuracy of pedestrian bounding boxes in person search and large-scale person matching. A novel person search framework is proposed, which includes: (1) A multi-layer cascade heatmap mechanism (MCHM) is proposed, which aggregates pedestrian features by multi-layer heatmaps cascaded and improves the accuracy of the pedestrian bounding box by optimizating the offset between the center of the bounding box and the center point. (2) A learnable part-based pedestrian feature weight calculation module is proposed, which can learn the weight of the part according to the importance of the part-based feature instead of manually set hyperparameters. (3) A group feature correlation graph convolution network (GFCGCN) is proposed, which can calculate the similarity between group pedestrian features and provide a more accuracy end to end person search work. Some ablation studies and comparative experiments on datasets CUHK-SYSU, PRW show that our model can effectively achieve more accuracy pearch search with accuracy of 88.7% rank-1 and 78.2% mAP.

Keywords: Pedestrian re-identification · Convolutional neural network · Graph convolutional network · Person detection · Person search

1 Introduction

Pedestrian re-identification is a research, which aims at matching a target person with a gallery of images. Although numerous methods have been proposed, most methods [1,2] mainly rely on external object detectors to detect pedestrians in the video, and then perform pedestrian matching on candidate sets. These methods treat detection and re-identification as two separate tasks. Existing pedestrian detectors inevitably produce false detections, missing detections, and misalignments, which will harm the final searching performance significantly. Person search has recently emerged as the task of finding a person, provided as a cropped exemplar, in a gallery of non-cropped images [3,4].

Y. Peng et al. (Eds.): PRCV 2020, LNCS 12306, pp. 242–254, 2020.
https://doi.org/10.1007/978-3-030-60639-8_21

Although the existing works have tried to address these bottlenecks, these works generally have the following deficiencies: 1) They ignored the impact of inaccurate pedestrian bounding boxes on pedestrian re-identification. The general object detection model usually detects multiple categories of objects. Obviously the wrong category will have a negative impact on pedestrian re-identification. 2) The simple pedestrian feature similarity measurement method is not robust enough. Both Euclidean distance and Consine distance require the feature to be highly robust. The information such as background texture has a great influence on distance measurement. 3) They do not consider the influence of the crowd features around the target pedestrian on the re-identification results. If the pedestrians around the target all appear in another camera, the confidence that the target appears in that camera should be higher.

To address this deficiency, an novel person search framework is proposed in this paper. This framework uses MCHM to aggregate information on features to reduce the impact of background texture in pedestrian features and improve the accuracy of pedestrian bounding boxes. And the GFCGCN module is proposed to calculate group pedestrian features so as to achieve more accurate pedestrian re-identification. The contributions of our model are as follows:

(1) A multi-layer cascade heatmap mechanism (MCHM) is proposed, which aggregates pedestrian features by multi-layer heatmaps and improves the accuracy of the pedestrian bounding box by optimizating the offset between the center of the bounding box and the center point.
(2) A learnable part-based pedestrian feature weight calculation module is proposed, which can learn the weight of the part according to the importance of the part-based feature instead of manually set hyperparameters.
(3) A group feature correlation graph convolution network (GFCGCN) is proposed, which can calculate the similarity between group pedestrian features and provide a more accuracy end to end person search work.

2 Related Work

In this section we first introduce prior art on the two separate tasks of person detection and person re-identification, and then introduce the person search.

2.1 Pedestrian Detection and Re-identification

In recent years, convolutional neural networks (CNNs) at pedestrian detection joint learning the classification model and the features in an end-to-end fashion [5]. Commonly used pedestrian detection models can be divided into single-stage models and two-stage models. While single-stage object detectors [6,7] are preferable for runtime performance, the two-stage strategy of Faster R-CNN remains the more robust general solution [8], versatile to tailor region proposals to custom scene geometries [9] and to add multi-task branches [10,11].

Person re-identification aims to associate pedestrians over non-overlapping cameras. Most previous methods try to address this task on two directions, i.e., feature representation and distance metric learning. Some methods design different kinds of hand-crafted features to achieve certain success on small datasets. But these methods are limited for large-scale searching. While there are two main trends in the modern CNN model learning: (1) by Siamese networks and contrastive losses; (2) by ID classification with crossentropy losses. In the first, pairs [12,13], triplets [14] or quadruplets [15] are used to learn a corresponding number of Siamese networks, by pushing or pulling the same or the different person ids, respectively. In the second, [16] define as many classes as people IDs, train classifiers with a cross-entropy loss, and take the network features as the embedding metric during inference.

2.2 Person Search

Person search is a recently introduced problem of matching a probe person bounding box against a set of gallery whole scene images [17]. Some methods [11,26] design online learning object functions to learn large number of identities in the training set and achieve great performance on recent person search datasets. However, these methods only employ individual appearance for verification, which ignores the underlying relationship between individuals in the scene. This is challenging due to the uncontrolled false alarms, misdetections, and misalignment emerging in the auto-detection process. The multi-scale matching problem turns out a more severe challenge in person search.

3 Method

In this section, the proposed person search framework will be introduced in detail. Firstly, a backbone network is used to encode features and a multi-layer cascaded heatmap mechanism (MCHM) is proposed to makes the bounding box more accurate, where the center point and bounding box are continuously optimized by training. Secondly, a group feature correlation graph convolution network (GFCGCN) is applied to output the similarity estimation. The overall structure of the framework is shown in Fig. 1.

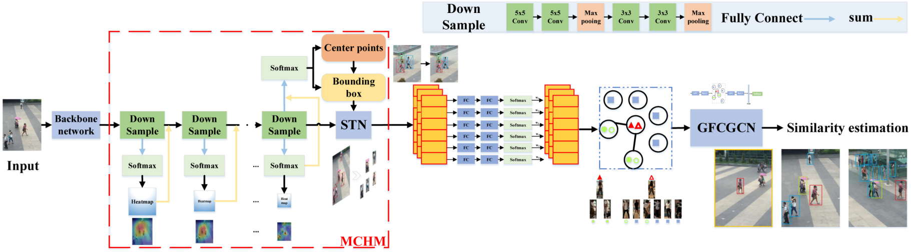
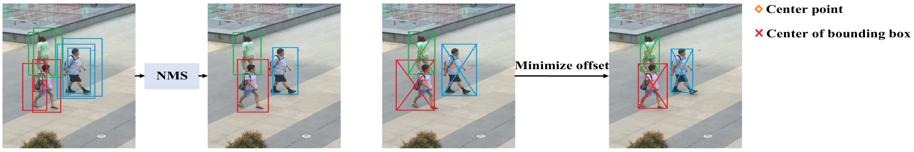


Fig. 1. An overview of the person search framework.

3.1 Multi-layer Cascade Heatmap and Anchor-Free Based Detection

The ResNet is adopted as backbone to extract deep features and then a multi-layer cascaded heatmap mechanism (MCHM) is proposed for further information aggregation. In a non-cropped image, it usually contains rich background information, which brings difficulty to detection and re-identification. To reduce the interference caused by background information, the MCHM outputs heatmaps from different levels of features to make the model pay more attention to pedestrians. Finally, multi-level heatmaps use up-sampling layers to aggregate information on features to achieve more accurate pedestrian detection.

The common object detection model obtains the bounding box coordinates by regression, which depends on the quality of the regression. However, this method usually causes information loss or contains too much noise information such as contains too much background information. In this case, the anchor-free based detect head of the MCHM not only outputs the bounding box coordinates, but also outputs the center point of the pedestrian. During the training steps, the bounding box is continuously corrected by minimizing the offset between the center point and the center of bounding box as shown in Fig. 2. Hence the center point is responsible for localizing the objects more precisely. Note that the benefits for pedestrian detection performance may be marginal. But it is critical for pedestrian re-identification because the pedestrian features extracted according to bounding box.



(a) Common object detection bounding box regression processing.

(b) Bounding box processing by optimizing center point offset in MCHM.

Fig. 2. An example diagram of pedestrian bounding box optimization process.

For each box $b^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ in the image, assign the object center (c_x^i, c_y^i) as $c_x^i = \frac{x_1^i + x_2^i}{2}$ and $c_y^i = \frac{y_1^i + y_2^i}{2}$, respectively. Then its location is obtained by dividing the stride $(\tilde{c_x^i}, \tilde{c_y^i}) = (\lfloor \frac{c_x^i}{4} \rfloor, \lfloor \frac{c_y^i}{4} \rfloor)$. Then the heatmap response at the location (x, y) is computed as $M_{xy} = \sum_{i=1}^{N} exp^{-\frac{(x - \tilde{c_x^i})^2 + (y - \tilde{c_y^i})^2}{2\sigma_c^2}}$. Where N denotes the number of objects in the image and σ_c denotes the standard deviation. The loss function is defined as pixel-wise logistic regression with focal loss:

$$L_{heatmap} = -\frac{1}{N} \sum_{xy} \begin{cases} (1 - \hat{M}_{xy})^{\alpha} log(\hat{M}_{xy}), & if \quad M_{xy} = 1; \\ (1 - \hat{M}_{xy})^{\beta} (\hat{M}_{xy})^{\alpha} log(1 - \hat{M}_{xy}) & otherwise \end{cases} \tag{1}$$

where $\hat{M}$ is the estimated heatmap, and α, β are the parameters.

Assume the outputs of the bounding box size and the offset as $\hat{S} \in R^{W*H*2}$ and $\hat{O} = R^{W*H*2}$, respectively. For each GT box $b^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ in the image, we can compute its size as $s^i = (x_2^i - x_1^i, y_2^i - y_1^i)$. Similarly, the GT offset can be computed as $o^i = (\frac{c_x^i}{4}, \frac{c_y^i}{4}) - (\lfloor \frac{c_x^i}{4} \rfloor, \lfloor \frac{c_y^i}{4} \rfloor)$. Denote the estimated size and offset at the corresponding location as $\hat{S}^i$ and $\hat{O}^i$, respectively. Then we enforce l_1 loss for the two outputs:

$$L_{box} = \sum_{i=1}^{N} ||o^i - \hat{o}^i||_1 + ||s^i - \hat{s}^i||_1 \tag{2}$$

3.2 Pedestrian Feature Extraction and Group Feature Proposal

Once the bounding box obtained, the pedestrians can be extracted by a STN module. Because there is a gap between feature coordinates and image coordinates. The STN module can correct this gap by affine transformation. As individual features are not sufficient for real world retrieval task, a group features is employed to help calculate the weights of the part-based features. Suppose the set of persons which appear on both probe and gallery scenes as positive feature pairs. The way of judging whether two features belong to the same person is to compute the similarity between the feature pairs. x_i^r, x_j^r is denoted as the $r-th$ part from feature i and j. As shown in Fig. 3, consider different feature parts, the final similarity $s(i,j)$ can be represented as the summation of different parts:

$$s(i,j) = \sum_{r=1}^{R} *w_r dist(x_i^r, x_j^r) \tag{3}$$

where dist denotes the Euclidean distance between x_i^r, x_j^r, R is the number of part (R = 6 in our framework). w_r is the contribution weight of the $r-th$ feature part.

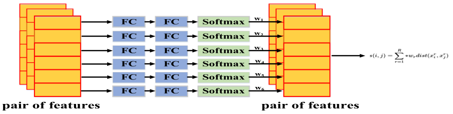


Fig. 3. A diagram of pair of features similarity weights calculation.

Because the weights of different parts are significantly different across samples, due to possible occlusions, different viewpoints and lighting conditions. The weight w_r will have a impact on the final similarity. In this case, the model uses two fully connected layers and a Softmax layer to output a learnable weights w_r. It takes in R pairs of feature vectors, and the Softmax layer output R normalized weights. Given an object pair (i,j), the corresponding label $y = 1$ if these two samples belong to the same person, otherwise $y = 1$, the loss function is as follows:

$$L = \begin{cases} 1 - s(i,j) & if \quad y = 1 \\ max(0, s(i,j) + \alpha) & if \quad y = -1 \end{cases} \tag{4}$$

This loss term builds a margin α between positive and negative pairs, and thus safeguards the discrminativeness of the embedded features.

3.3 Group Feature Correlation Graph Learning

For a given image pair A, B. The motivation is to determine whether the target in image A also appears in image B. Therefore, assuming that a target is captured in image A and B, respectively. All need to be done is to determine whether these two targets are the same pedestrian. For probability events, if most of the pedestrians around the target in A appear in B, then there is a higher confidence that the target also appears in image B. Based on this, we fully consider the impact of crowd on pedestrian re-identification.

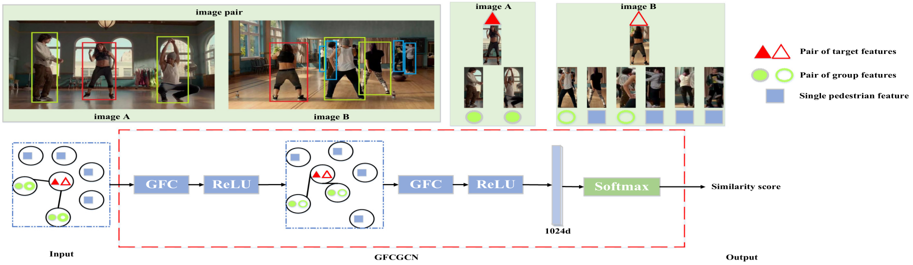


Fig. 4. An overview of the model structure of the GFCGCN network.

Given K pedestrian group feature pairs $(A_i, B_i), i \in 1, .., K$. A graph is dessigned to jointly take the target pairs and the K group feature pairs as well as single features (only appear in one image) into consideration as shown in Fig. 4. In this graph, the target pedestrian node is the center of the graph, which is connected to all the group feature nodes for information aggregation and node weight updation.

Assume the graph mentioned above is denoted by G. Where $G = (V, E)$, V represents the N-dimensional feature vector, and E represents the edge set of the graph. Each node is assigned with a pair of features $(X_{A_j}, X_{B_j}), j \in 0, ..., K$. Suppose the images have K group feature pairs, then $N = K + 1$. We define $X \in R^{N*2d}$ and $A \in R^{N*N}$, where d is the pedestrian feature dimension. A denote the adjacent matrix associated with graph G and it can be expressed by the following formula:

$$A_{i,j} = \begin{cases} 1 \quad if \quad i = 1 \quad or \quad j = 1 \quad or \quad i = j, \\ 0 \quad otherwise \end{cases} \tag{5}$$

where $i, j \in 1, ..., N$. To better implement the model, the adjacency matrix A is normalized for the ease of learning. The adjacency matrix A can be seen as a stack of $\{A_1, ..., A_T\}$, each A_t is normalized symmetrically by $A_t = \Lambda_t^{-\frac{1}{2}} * \hat{A}_t * \Lambda_t^{-\frac{1}{2}}$. Where $\hat{A}_t = A_t + I$ and Λ_t is the diagonal node degree matrix of $\hat{A}_t$. $\hat{A}$ and Λ are used to denote the stack of $\hat{A}_t$ and Λ_t, respectively. Finally, a group feature correlation graph convolution network (GFCGCN) is proposed to update the weights and output the similarity. The network structure can be shown in Fig. 4 and the layer-wise GFCGCN propagates as follows:

$$GFC(V^l, A)^{l+1} = \sigma(\Lambda^{-\frac{1}{2}} * \hat{A} * \Lambda^{-\frac{1}{2}} * V^{(l)} * W^{(l)})) \tag{6}$$

where $V^{(l)}$ is the outputs of the $l-th$ layer, and $V^{(0)} = X$ as input. $W^{(l)}$ is the learnable parameters and σ is the ReLU activation function. Finally, a fully connected layer is applied to merge all the vertices into a 1024-dimensional feature vector. And a binary Softmax layer is employed supervise network training.

4 Experiments and Analysis

4.1 Datasets

CUHK-SYSU. The CUHK-SYSU dataset [17] consists of 18184 images, labeled with 8,432 identities and 96,143 pedestrian bounding boxes (23,430 boxes are ID labeled). The images, captured in urban areas by hand-held cameras or from movie snapshots, vary largely in viewpoint, lightning, occlusion and background conditions.

PRW. The PRW dataset [4], acquired in a university campus from six cameras, consists of 11,816 images with 43,110 bounding boxes (34,304 boxes are ID labeled) and 932 identities. Compared to CUHK-SYSU, PRW is with features less images and IDs but many more bounding boxes per ID (36.8, against 2.8 in CUHK-SYSU), which makes it more challenging.

4.2 Implementation Details

An ImageNet pretrained ResNet-50 model is applied as a backbone. The model is trained 60 epochs with the Adam optimizer and a starting learning rate of 0.001. The learning rate is reduced by 10% every 10 epochs. All the training images are resized to 512 $*$ 128. Besides, a standard data augmentation including rotation, scaling and color jittering is applied to enhance data. The model is implemented on Pytorch, trained and tested on two Tesla P100 GPUs.

4.3 Multi-layer Cascade Heatmap Mechanism

The MCHM is used to detect pedestrians and extract features from non-cropped images. In order to verify the effectiveness of the MCHM, some relevant comparative experiments and ablation experiments are conducted.

Table 1. A comparison of accuracy between the proposed MCHM and common object detection methods.

Methods	AP	AP_{50}	AP_{60}
Faster RCNN [19]	26.8	46.7	36.7
RGB-D Faster RCNN [18]	36.7	59.5	38.9
MFI-SSD [20]	45.0	63.5	46.7
CornerNet [21]	47.6	63.7	53.1
CenterNet [22]	52.4	64.8	56.3
MCHM (ours)	**56.8**	**70.1**	**57.1**

Comparative Experiments. Some common object detection models propose bounding boxes by anchor, which can be summarized as anchor-based methods. Besides, there are some key point-based methods called anchor-free methods. In this section, the proposed MCHM is compared with two sorts of common object detection methods on the datasets CUHK-SYSU. Among them, RGB-D Faster RCNN [18], Faster RCNN [19] and MFI-SSD [20] are employed as the anchor-based method. And CornerNet [21] and CenterNet [22] are employed as the anchor-free method. The results are shown in Table 1. It can be seen from the experiment that the proposed MCHM is significantly better than the common object detection model under the optimization of the pedestrian center point.

Ablation Study. Different level of cascaded heatmap layers (3, 4, 5 layers) are applied to the ablation experiment to explore the performance of the MCHM. The experimental results are shown Fig. 5.

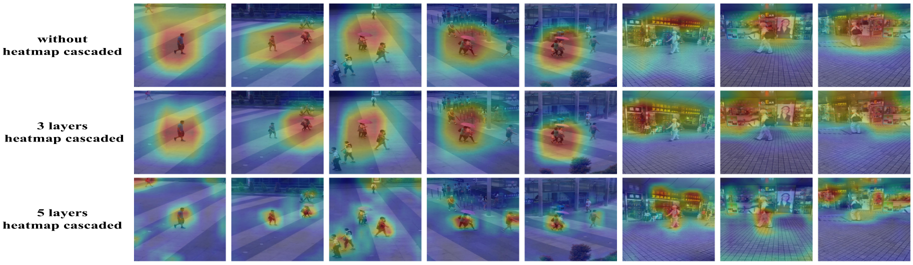


Fig. 5. The effect of MCHM on pedestrian feature extraction. The greater the number of cascaded heatmaps, the more clustered pedestrian features are and the less background information it contains.

Without using MCHM, the model can only learn some roughly information around pedestrian features. Usually that features contain too much background information, which is harmful to pedestrian re-identification. The MCHM can

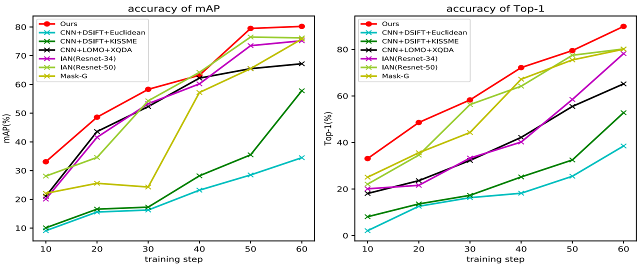


Fig. 6. The accuracy of the comparative experiments on performance.

effectively reduce the background information. In addition, the different level of cascaded heatmaps also has a greater impact on the information aggregation. Compared with the 3-layer heatmaps cascaded, The 5-layer structure can more accurately pay attention to the pedestrians. Overall, the MCHM can effectively aggregate pedestrian information.

4.4 Group Feature Correlation Graph Learning

Conventionally Euclidean distance is used to calculate the similarity between features. However, this approach often requires the features to be very robust. The quality of the features affects the similarity result directly. This model combines target pedestrians and contextual pedestrians to form group features to calculate the similarity between features. The proposed GFCGCN model measures the correlation between group features and finally makes a comprehensive similarity estimation. Similarly, some research is carried out to explore the effectiveness of the model. Firstly, some comparative experiments are performed between the GFCGCN and some existing metric learning methods. Secondly, some ablation experiments are conducted to study the influence of this part on the final results. Finally, some person search results of this framework is demonstrated (Fig. 6).

Comparative Experiments. The model is compared with some previous metric learning re-identification models such as IAN [11], Dis-GCN [25], as well as some other hand-crafted features such as DSIFT [23] and LOMO [24]. The experimental quantification results are shown in Table 2.

Ablation Study. In this subsection, the MCHM and GFCGCN are combined to do ablation experiments to explore the person search result. The ablation experiments use different backbone networks, GFCGCN and conventional distance formula. The quantitative results are described in Table 3. In addition, the influence of the number of group features K can be qualitatived in Fig. 7 and Fig. 8. It can be seen from the curve in the figure that for five pedestrians appearing at the same time, this model can significantly improve the effect of pedestrian search. Therefore, this framework can be more suitable for person search in crowded scenes.

Table 2. Quantitative results of some comparative experiments.

Methods	Datasets	mAP (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
CNN+DSIFT+Euclidean [23]	CUHK-SYSU	34.5	38.6	45.8	56.8
CNN+DSIFT+KISSME [23]	CUHK-SYSU	47.8	53.5	60.8	78.9
CNN+LOMO+XQDA [24]	CUHK-SYSU	68.9	74.1	80.2	89.9
IAN(Resnet-34)IAN [11]	CUHK-SYSU	73.1	78.0	84.2	92.4
IAN(Resnet-50)IAN [11]	CUHK-SYSU	75.0	80.5	86.0	96.3
Dis-GCN [25]	CUHK-SYSU	75.8	80.1	90.2	93.2
Ours	CUHK-SYSU	**78.2**	**88.7**	**94.8**	**96.5**
CNN+DSIFT+Euclidean [23]	PRW	17.4	23.8	32.3	40.7
CNN+DSIFT+KISSME [23]	PRW	18.4	25.8	30.1	40.2
CNN+LOMO+XQDA [24]	PRW	20.4	23.1	34.8	43.8
IAN(Resnet-34)IAN [11]	PRW	23.0	50.8	60.8	74.7
IAN(Resnet-50)IAN [11]	PRW	35.8	56.7	65.3	75.8
Dis-GCN [25]	PRW	40.5	56.8	62.4	70.0
Ours	PRW	**57.8**	**72.3**	**80.5**	**86.4**

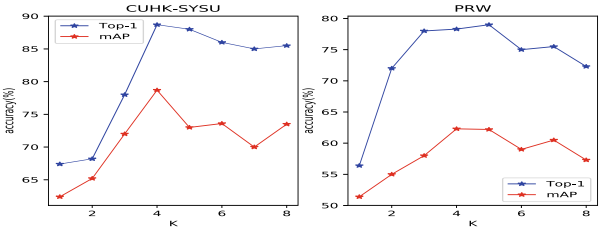


Fig. 7. The impact of group feature size K on performance.

Table 3. An ablation study of the proposed MCHM mechanism on dataset CUHK-SYSU.

Model structure	Similarity estimation	mAP (%)	Top-1 (%)
Resnet-34+MCHM	**GFCGCN**	**73**	**78.3**
Resnet-34+MCHM	Euclidean distance	58.1	63
Resnet-34+MCHM	cosine distance	56.3	59.8
Resnet-50+MCHM	**GFCGCN**	**78.2**	**88.7**
Resnet-50+MCHM	Euclidean distance	68.4	71.6
Resnet-50+MCHM	Cosine Distance	65.3	70.9

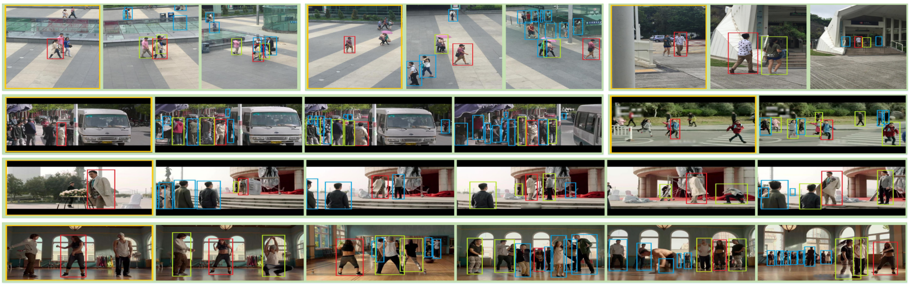


Fig. 8. The performance of different crowd sizes in the experiment. The red bounding box represents the target pedestrian, the green bounding box represents the surrounding crowd, and the blue bounding box represents the pedestrian first appeared in the scene. (Color figure online)

5 Conclusion

In this work, a novel person search framework with a MCHM module and GFCGCN module is proposed. The framework combines pedestrian detection and re-identification as one task and significantly improves the person search result. Instead of identifying target independently, the framework combines the surrounding crowds to form group features for re-identification. The framework has been verified on public datasets and achieved better re-identification results. It can be used to implement an end-to-end person search work in the surveillance system.

Acknowledgments. This research is supported by National Natural Science Foundation of China (61972183, 61672268) and National Engineering Laboratory Director Foundation of Big Data Application for Social Security Risk Perception and Prevention.

References

1. Wu, D., Zhang, K., Zheng, S.J., et al.: Random occlusion recovery for person re-identification. J. Imaging Sci. Technol. **63**(3), 30405-1–30405-9 (2019)
2. Wu, Q., Dai, P., Chen, P., et al.: Deep adversarial data augmentation with attribute guided for person re-identification. Signal Image Video Process. 1–8 (2019). https://doi.org/10.1007/s11760-019-01523-3
3. Liu, H., Feng, J., Jie, Z., et al.: Neural person search machines. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 493–501 (2017)
4. Zheng, L., Zhang, H., Sun, S., et al.: Person re-identification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1367–1376 (2017)
5. Guo, S., Bai, Q., Zhou, X.: Foreign object detection of transmission lines based on faster R-CNN. In: Kim, K.J., Kim, H.-Y. (eds.) Information Science and Applications. LNEE, vol. 621, pp. 269–275. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-1465-4_28

6. Lin, T.Y., Goyal, P., Girshick, R., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
7. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
8. Durkee, M.S., Sibley, A., Ai, J., et al.: Improved instance segmentation of immune cells in human lupus nephritis biopsies with Mask R-CNN. In: Medical Imaging 2020: Digital Pathology, vol. 11320, p. 1132019. International Society for Optics and Photonics (2020)
9. Jiang, H., Li, S., Liu, W., et al.: Geometry-aware cell detection with deep learning. MSystems **5**(1) (2020)
10. Hasan, I., Tsesmelis, T., Galasso, F., et al.: Tiny head pose classification by bodily cues. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2662–2666. IEEE (2017)
11. Xiao, J., Xie, Y., Tillo, T., et al.: IAN: the individual aggregation network for person search. Pattern Recogn. **87**, 332–340 (2019)
12. Jiang, M., Li, C., Kong, J., et al.: Cross-level reinforced attention network for person re-identification. J. Vis. Commun. Image Represent. 102775 (2020)
13. Şerbetçi, A., Akgül, Y.S.: End-to-end training of CNN ensembles for person re-identification. Pattern Recognit. 107319 (2020)
14. Zhao, C., Lv, X., Zhang, Z., et al.: Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. IEEE Trans. Multimedia (2020)
15. Zhang, C., Yue, J., Qin, Q.: Deep quadruplet network for hyperspectral image classification with a small number of samples. Remote Sens. **12**(4), 647 (2020)
16. Ye, M., Shen, J., Lin, G., et al.: Deep Learning for Person Re-identification: A Survey and Outlook. arXiv preprint arXiv:2001.04193 (2020)
17. Xiao, T., Li, S., Wang, B., et al.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3415–3424 (2017)
18. Zhu, X., Chen, C., Zheng, B., et al.: Automatic recognition of lactating sow postures by refined two-stream RGB-D faster R-CNN. Biosyst. Eng. **189**, 116–132 (2020)
19. Mai, X., Zhang, H., Jia, X., et al.: Faster R-CNN with classifier fusion for automatic detection of small fruits. IEEE Trans. Autom. Sci. Eng. (2020)
20. Zhou, J., Chen, B., Zhang, J., et al.: Multi-scales feature integration single shot multi-box detector on small object detection. In: MIPPR 2019: Pattern Recognition and Computer Vision, vol. 11430, p. 114300E. International Society for Optics and Photonics (2020)
21. Law, H., Deng, J.: Cornernet: detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
22. Duan, K., Bai, S., Xie, L., et al.: Centernet: keypoint triplets for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6569–6578 (2019)
23. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR, pp. 3586–3593 (2013)
24. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR, pp. 2288–2295 (2012)

25. Ktena, S.I., et al.: Distance metric learning using graph convolutional networks: application to functional brain networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 469–477. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_54
26. He, Z., Zhang, L.: End-to-end detection and re-identification integrated net for person search. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11362, pp. 349–364. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20890-5_23