A Quantitative Analysis of the Effect of Human Detection and Segmentation Quality in Person Re-identification Performance

Thuy-Binh Nguyen *†, Hong-Quan Nguyen *¶, Thi-Lan Le *, Thi Thanh Thuy Pham *‡, Ngoc-Nam Pham §
* Computer Vision Department, MICA International Research Institute,
Hanoi University of Science and Technology, Vietnam
Email: Thi-Lan.Le@mica.edu.vn
†Faculty of Electrical and Electronics Engineering, University of Transport and Communications, Hanoi, Vietnam
‡Faculty of Security and Information Technology, Academy of People Security, Hanoi, Vietnam
§VinUniversity Project, Hanoi, Vietnam
¶Faculty of Information Technology, Viet-Hung Industrial University, Hanoi, Vietnam

***Abstract*—Person re-identification, a problem of person identity association across camera views at different locations and times, is the second step in two-steps system for automatic video surveillance: person detection, tracking and person re-identification. However, most of the reported person Re-ID methods deal with the human regions of interest (ROIs) which are extracted manually with well-aligned bounding boxes. They mainly focus on designing discriminative feature descriptors and relevant metric learning on these manually-cropped human ROIs. This paper aims at answering two questions: (1) Do human detection and segmentation affect the performance of person re-identification?; (2) How to overcome the effect of human detection and segmentation with the state of the art method for person re-identification? To answer these two question, quantitative evaluations have been performed for both single-shot and multi-shot scenarios of person re-identification. Different state-of-the-art methods for human detection and segmentation have been evaluated on two benchmark datasets (VIPeR and PRID2011). The obtained results allow to give some suggestions for developing fully automatic video surveillance systems.**

I. Introduction

Person re-identification (ReID) is defined as a task of association of multiple appearance images of a pedestrian when he/she moves in a non-overlapping camera network. It is getting increasing attention in the computer vision and recognition community, with many applications in robotics, control and person retrieval systems, etc.

In fact, most of the proposed works on person ReID only deal with human regions of interest (ROIs) which are extracted manually with well-aligned bounding boxes [1], [2]. These works can be roughly classified into two main approaches, such as building a discriminative descriptor for person representation and learning an effective metric distance for person matching. Therefore, the performance of person ReID is only evaluated from this. However, relating to person ReID efficiency, especially for the real applications, other components should be considered, such as person detection and tracking. They are two important and prerequisite steps of a complete person ReID system. In this paper, the influence of human detection followed by segmentation step on person ReID is evaluated. Based on this, some valuable recommendations are given out for building a full system of person ReID.

II. Related work

In this section, the related works on a fully-automated system of person ReID is reviewed. This system contains not only person ReID phase but also other crucial phases of human detection and tracking. In fact, there are a few works focusing on this in contrast to a wide range of other ones for the only phase of person ReID. In [3], the authors proposed a fully-automated person Re-ID system in the real scenarios of non-overlapping camera network. In this work, performance of person ReID is evaluated on both phases of human detection and person ReID. It was proven from this work that automatic human detection, which is the first and obvious phase of a fully-automated person Re-ID system put more burdens on person ReID in comparison with the other popular ones of manual detection. Therefore, together with improving the efficiency of person ReID, the authors proposed to improve the performance of human detection by applying an effective shadow removal method. Other related work [4] also interested in evaluating the entire system of person ReID on surveillance data captured from multiple cameras. These cameras have non-overlapping FOVs covering a wide area of moving paths. In this proposal, some advanced techniques of auto human detection, tracking and ReID are applied, such as DPM (Deformable Part-Based Model) and HOG (Histogram of Oriented Gradients) for detection, Tracking-by-Detection for tracking and gait feature for ReID. Zheng et al [5] introduced comprehensive baselines for end-to-end person ReID in raw video frames. A novel dataset is provided in this work, called Person Re-identification in the Wild (PRW), and extensive experiments are conducted by combining various detectors and recognizers to improve the overall person re-identification performance. In [6], the authors claimed that if background removal is performed directly by applying binary

978-1-7281-1829-1/19/$31.00 ©2019 IEEE

masks which might cause loss of information and results in a slightly worse performance compared to case of using original images. Therefore, they introduced a proposal in which background is removed in the feature-level extraction. It is called as Mask-Guided Contrastive Attention Model (MGCAM) with a binary mask considered as an additional input which is accompanied by an RGB image to enhance feature learning. The effectiveness of this method is proven by impressive results on several public datasets. In our work, towards to a complete person ReID system, we consider other crucial phases that are person detection and segmentation. The effect of person detection and segmentation on person ReID performance is carefully evaluated on both single shot and multishot scenarios.

III. Proposed Method

A. Overall framework

Figure 1 shows the framework of a fully automatic person ReID system. It contains four main steps: person detection, segmentation, tracking and person ReID. While person detection step aims at determining the person region (bounding box) in images captured from surveillance cameras, person segmentation is used for removing background from person bounding box. Then, person bounding boxes within a camera field of view (FoV) are associated through person tracking step. Finally, person ReID aims to associate instances of the same person when he/she moves from on camera FoV to the others ones. It is worth to note that in some systems, person segmentation and person detection are coupled. In this work, in order to understand in detail the affection of person detection and segmentation on the person ReID performance, we perform both person detection and segmentation. Obviously, tracking step also plays an important role in assuring person ReID performance. However, the evaluation on the effect of tracking on person ReID is out of the scope of this paper.

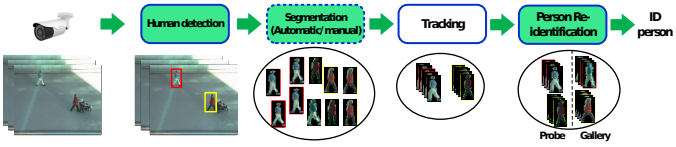


Fig. 1. The proposed framework for a fully automatic person ReID system.

B. Person detection and segmentation

Pedestrian detection is a prerequisite step in a person ReID system. There are many detectors are proposed. Each of them has its own advantages and disadvantages in the metrics of detection accuracy or computational speed. For a person ReID system, these methods together with segmentation techniques should be evaluated to show their impact on person ReID. Based on this evaluation, some suggestions for building a complete person ReID system can be given.

In this paper, concenring person detection, we employ three state-of-the-art person detection techniques that are Aggregate Channel Features (ACF) [7], You Only Look One (YOLO) [8], and Mask R-CNN [9]. ACF detector based on feature extraction in which feature vectors are extracted and aggregated on multi-layer scales and Adaboost classifier to predict object regions of interest. YOLO is an object detection algorithm that is much different from the region based algorithms as Faster R-CNN [10]. For YOLO detector, a single network is employed to predict the bounding boxes and the probability that each box belongs to different classes. We propose to use YOLO v3 because of the balance between the speed and the accuracy. Concerning person segmentation, Pedparsing [11] method is used thanks to its effectiveness for cropped images from the result of the detectors that mentioned above. Another way to perform simultaneously person detection and segmentation is use Mask R-CNN [9]. This network is built by adding two more convolutional layers to generate a mask for a corresponding bounding box. It is worth to note that, we used pre-trained model on COCO dataset for both YOLO and Mask-RCNN. Figure 2 shows an example of person detection and segmentation obtained by these above-mentioned methods.

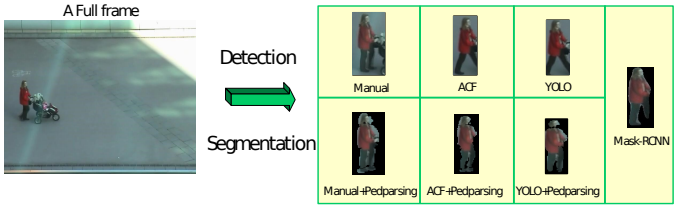


Fig. 2. An example for automatic person detection and segmentation results on PRID 2011 dataset.

C. Person Re-identification

Depending on the number of images used for person representation, person ReID can be divided into single-shot and multi-shot. In single-shot person ReID, each person has one sole image while multiple images are available in multi-shot person ReID. Single-shot approach does not seems to be suitable for a practical application, however, obtained results on this approach would be scalable for multi-shot one. Most studies on person ReID are interested in two crucial issues that are feature extraction and metric learning. The purpose of feature extraction is to build a descriptor is not only discriminative but also robust to strong variations in illumination, viewpoints, poses, etc. While metric learning aims at minimizing the distance between intra objects, inversely, maximizing the distance between extra objects. The process is understood that projecting extracted features onto a sub-space makes person matching step become simpler. In our work, we take one state of the art method for person ReID [12] that uses Gaussian of Gaussian (GOG) descriptor and Cross-view Quadratic Discriminant Analysis (XQDA) metric learning. This method outperforms a number of the state of the art methods for single-shot person ReID. In order to handle multi-shot problem, some works turn multi-shot problem into single-shot one by applying different pooling techniques, such as max-, min-, or average-pooling. Some others prefer to compare two sets of

feature vectors, namely set-to-set matching technique [13]–[15]. In our work, we survey both of these methods to assess their effectiveness for the person matching phase. For the first approach, average-pooling technique is exploited and for the other one, we consider Block Sparsity for Re-Identification (SRID) [16]. Average-pooling means to take the average value of all extracted feature vectors which are corresponding to all instance images of a given person. And, comparison between two objects is considered as comparison between two feature vectors. For the SRID method, the distance of two persons are calculated by summing all distances between each probe feature vector to the set of gallery features via a dictionary constructed by the gallery persons.

In addition, to understand the role of two main components in person ReID that are descriptor and person matching, we make a comparison between the chosen method with the others methods that used LOMO descriptor [17] for feature extraction and Cosine distance for person matching step.

IV. Experimental results

A. Dataset and evaluation metric

1) Dataset: In order to make a comprehensive evaluation, we conduct extensive experiments on both single-shot dataset (VIPeR) and multi-shot one (PRID 2011). Each dataset is split into two halves, one for training and one for test phase. This process is repeated randomly 10 times and the reported results are the average one from these.

VIPeR (Viewpoint Invariant Pedestrian Recognition) dataset [18] is considered as one of the most challenging datasets for single shot person ReID. This dataset contains 1,264 images of 632 persons observed by two static cameras. We follow the experiment set-up introduced in [17].

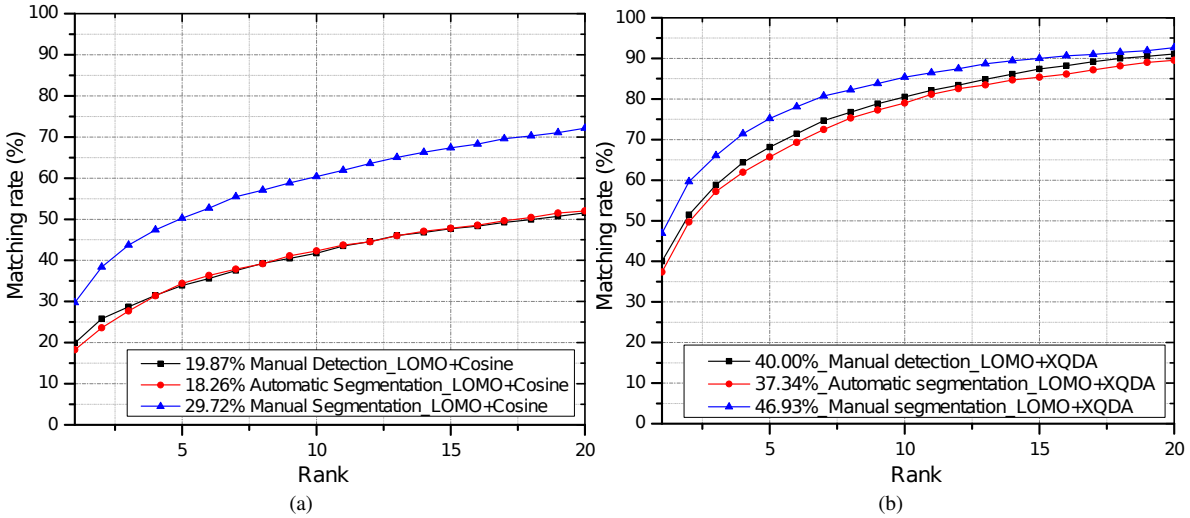
PRID 2011 (Person Re-ID) dataset includes 385 persons in view A, and 749 persons in view B. However, only 200 persons appear on both cameras. According to the experimental setting in [19], we only take 178 persons who have image sequences contain more than 21 images for our work.

2) Evaluation metric: In general, Cumulative matching characteristic (CMC) curve are used to estimate the performance of a proposal method for person ReID. This curve show the probability that the correct matching rates for a given query person within the top K ranks.

B. Experimental results

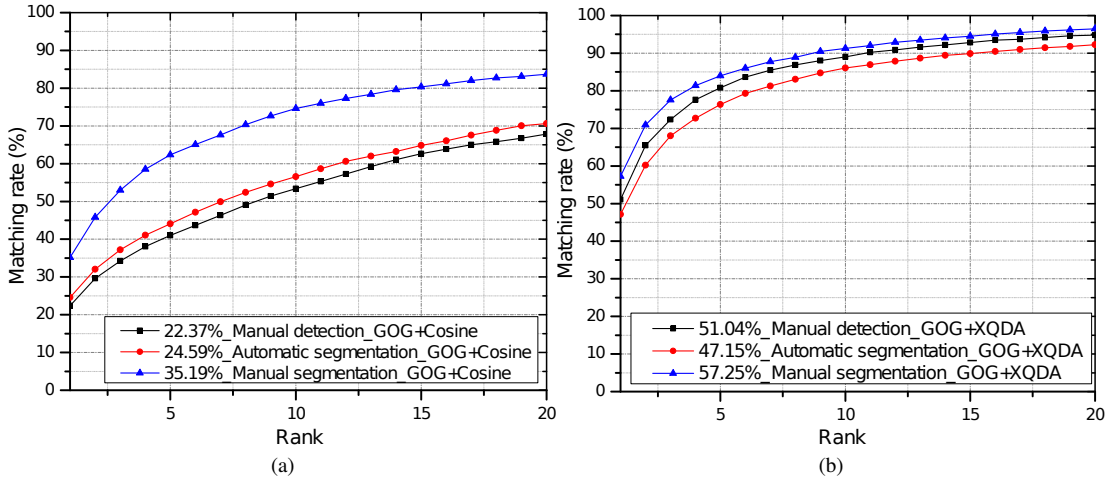
1) Evaluation for single shot scenario: As full frame in VIPER dataset is not available, in this experiment, we can only evaluate the effect of person segmentation. Two methods of person segmentation are considered: manual segmentation via Interactive Segmentation Tool and automatic segmentation based on Pedparsing method as described in section III.B. The obtained results on LOMO features with two distance metrics that are Cosine and XQDA and those on GOG features are shown in Fig. 3 and Fig. 4, respectively. Several observations can be given as follows. First, in four cases, manual segmentation obtains the best results. The manual segmentation allows to obtain more than 10% and 6% of improvement over manual detection with Cosine distance and XQDA on both LOMO and GOG features. This means that background plays an important role in person ReID performance. Second, automatic segmentation and manual detection obtains similar results while using Cosine distance on both LOMO and GOG feature. These results can be explained that automatic segmentation allows to remove background information. However, it also remove person information especially silhouette of the persons that is crucial information for distinguishing persons. However, it is interesting to see that while using XQDA, the manual detection outperforms automatic segmentation with approximately 3% of improvement. The reason is that XQDA allows to learn the relation between cross-view cameras. In this case, it learns also the relations of background informations. Therefore, the performance obtained with XQDA is much better than with Cosine distance. Finally, the chosen method for person ReID with two components that are GOG feature and XQDA is effective for person ReID. GOG feature steps obtains up to 6% of improvement in comparison with LOMO feature while XQDA allow to increase the accuracy from 17.21% to 28.67% compared to Cosine distance.

Moreover, we make an additional experiment for single-shot approach on PRID 2011 dataset. A bounding box for each individual is chosen randomly. After that, our frame work is applied on these bounding boxes. Because of choosing only a random image for each person, the identities as well as the number of persons in this experiment are not similar to those in multi-shot case on PRID 2011, which are often used in the most existing works. In this experiment, for detection human stage, we utilize ACF, YOLO, and Mask-RCNN. And, for segmentation, Mask-RCNN and Pedparsing are applied. It is worth to note that Mask-RCNN is used for detection and segmentation purpose simultaneously. For feature extraction, we utilize GOG descriptor for person representation based on obtained results in Figures 3 and 4. Performance of the fully automatic system are evaluated on some cases indicated in Fig.5 with two cases (without/with segmentation). Observing this Figure, we can provide several conclusions as follow. First, when comparing corresponding curves on left and right side Figures, segmentation stage provides worse results compared to those in case of applying only detection. This result means that background removal with binary masks, which may a reason for information loss and smoothness of an image, is not an optimal choice to improve the performance of person ReID task. The matching rates at rank-1 when applying segmentation process are reduced by 10.9%, 8.53%, 10.33% when compared to those in case of manual, ACF, and YOLO methods, respectively. Moreover, in comparison between considered detectors, obtained results indicate that ACF detector achieves a better performance than YOLO one in both cases (without/with segmentation). The matching rates at rank-1 when using ACF detector are higher by 2.47% and 4.27% compared to YOLO detector in case of without and with segmentation, respectively. In addition, the effectiveness of ACF detector can achieve the performance of manual detection. One remarkable point is that Mask-RCNN provides



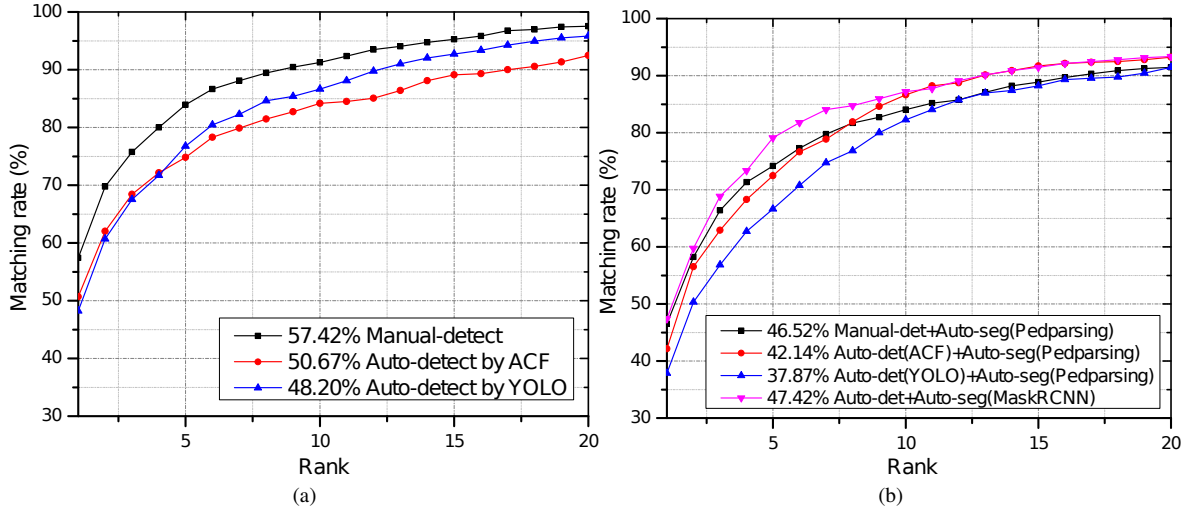
(a) (b)

Fig. 3. CMC curves of three evaluated scenarios on VIPER dataset with LOMO feature and (a) Cosine distance and (b) XQDA.



(a) (b)

Fig. 4. CMC curves of three evaluated scenarios on VIPER dataset with GOG feature and (a) Cosine distance and (b) XQDA.



(a) (b)

Fig. 5. CMC curves of three evaluated scenarios on PRID 2011 dataset in single-shot approach (a) Without segmentation and (b) with segmentation.

an impressive results that is competitive over manual detection. This bring a hopefulness for a fully automatic system to be practical.

2) Evaluation for multi-shot scenario: In the first experiment, we realize that XQDA technique is much more effective than Cosine distance. Therefore, for multi-shot scenario we only apply XQDA technique for person matching step. In the Section IIIC, we mention two approaches to solve multi-shot person re-identification. In the first experiment for multi-shot scenario, we would like to evaluate the effectiveness of the two mentioned approaches. Average-pooling and SRID [16] are applied on manual detected images for PRID 2011 dataset. The experiments are conducted on 10 different splits and the final result is the average value. As seen in Fig.6, we realize that the use of average pooling can achieve a better performance than SRID method [16]. Matching rate at rank-1 when applying average-pooling for person matching is 90.56% compared to 78.20% in SRID method. Noted that in this experiment, GOG is used for feature extraction phase.

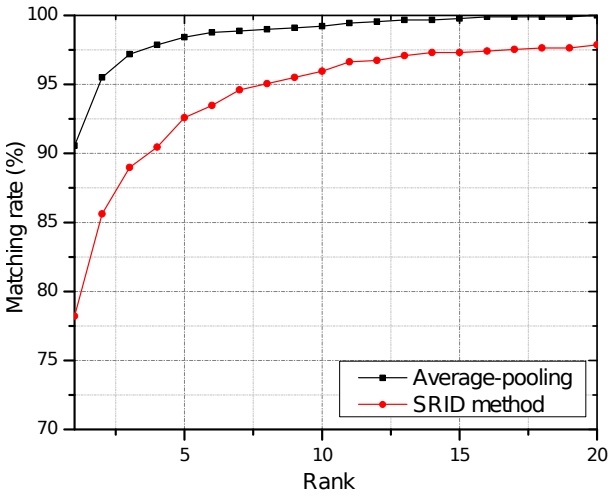


Fig. 6. Comparison between two techniques: average-pooling and SRID [16] on manual detected images for PRID 2011 dataset with GOG descriptor and XQDA for feature extraction and metric learning, respectively.

From above results, for the second experiment on multi-shot person re-identificaion, we only employ average-pooling technique for generating the final signature for each individual. This is really not only a simple but also effective strategy when turn multi-shot approach into single-shot one.

Figure 7 indicates the matching rates on PRID 2011 dataset when employing LOMO and GOG descriptors with XQDA technique with manual detection provided by the author dataset, one of the considered automatic detection techniques and automatic segmentation using Pedparsing after the automatic detection stage. It is worth to note that in order to make a comparison between automatic detection and manual detection, we keep only the person ROIs from automatic detection whose IoU with ROIs in manual detection is greater than 0.4. As seen in this Figure, GOG descriptor helps to increase the matching rates at rank-1 from 4.49% to 10% compared to LOMO descriptor. It is interesting to see that

TABLE I
COMPARISON RECOGNITION RATES WHEN USING EITHER LOMO OR GOG DESCRIPTOR AND XQDA FOR PERSON MATCHING ON PRID 2011.

Rank	Methods	Manual_detect	Auto_detect	Auto_detect+segment
R=1	LOMO	83.48	86.52	78.76
	GOG	**90.56**	**91.01**	**88.76**
R=5	LOMO	97.04	**99.35**	95.02
	GOG	**98.43**	98.43	**98.43**
R=10	LOMO	**99.68**	**100.00**	97.16
	GOG	99.21	99.33	**98.99**
R=20	LOMO	99.91	**100.00**	98.86
	GOG	**100.00**	99.89	**99.55**

the person re-identification results in case of using automatic detection is slightly better than those of manual detection. The reason is that automatic detection results very well-aligned bounding boxes while manual detection defines a larger bounding box for pedestrians. And, one more remarkable point is that when quality of person detection is relative good the segmentation step is not necessary. It can be shown in Fig. 7, the matching rates at rank-1 are reduced 7.76% and 2.25% on LOMO and GOG descriptor, respectively. It is an helpful recommendation for a fully automatic person ReID system. In order to show more clearly the comparison recognition rates two cases of using LOMO or GOG descriptors, we present these results in Table I (better results are in bold). Observing this Table, we realize that GOG descriptor outperforms LOMO in almost cases. Only in case of applying automatic detection, performance of LOMO is better than that of GOG in rank -5, -10, -20, however, the differences are very slightly (smaller than 1%). Finally, Table II shows the comparison of the proposed method with state of the art methods. From this Table, we realize that our work outperform over all mentioned even for deep learning approach [19], [20].

V. CONCLUSIONS AND FUTURE WORK

Based on obtained results we can confirm that the two previous steps affect on person ReID accuracy. However, the effect is much reduced thanks to the robustness of the descriptor and metric learning. The obtained results allow to give two suggestions. First, if automatic person detection step provide a relatively good performance, segmentation is not required. This helps to improve the computational time as segmentation step is time consuming. Second, multi-shot is preferred choice because this scenario considers all instances of one person. Therefore, it allows to remove poor detection results if they occur in few instances. In the future, we will evaluate the effect of person tracking in person ReID in order to give a complete recommendation for developing fully automatic surveillance systems.

ACKNOWLEDGEMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2017.12.

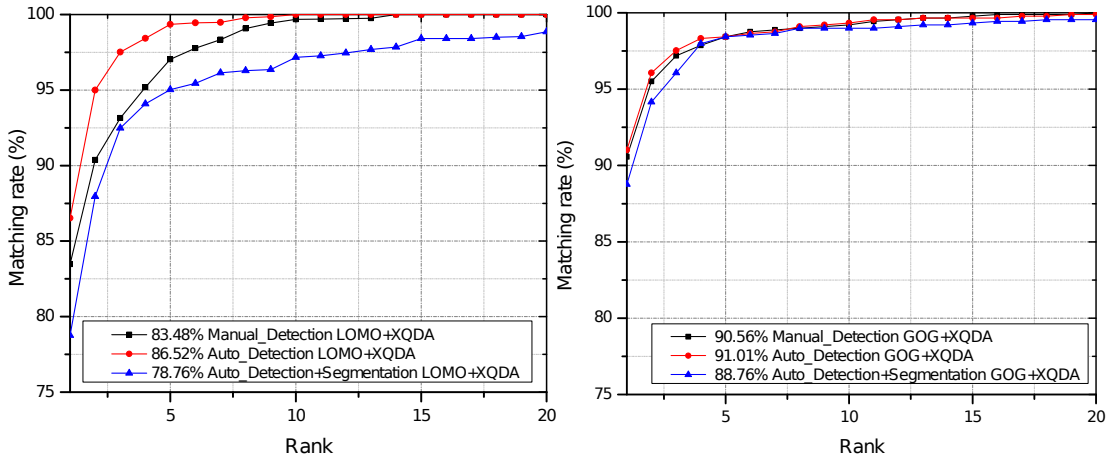


Fig. 7. CMC curves of three evaluated scenarios on PRID 2011 dataset when using XQDA technique with (a) LOMO descriptor and (b) GOG descriptor.

TABLE II
COMPARISON OF THE PROPOSED METHOD WITH STATE OF THE ART METHODS FOR PRID 2011 (THE TWO BEST RESULTS ARE IN BOLD).

Methods	**R 1**	**R 5**	**R 10**	**R 20**
HOG+DVR [21]	28.9	55.3	65.5	82.8
TAPR [20]	68.6	94.4	97.4	98.9
GOG+LSTM [19]	70.4	93.4	97.6	99.3
Our method with manual detection	**90.6**	**98.4**	**99.2**	**100**
with automatic detection	**91.0**	**98.4**	**99.3**	**99.9**
with automatic detection and segmentation	88.8	98.36	99.0	99.6

REFERENCES

[1] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1363–1372.

[2] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2018.

[3] T. T. T. Pham, T.-L. Le, H. Vu, T. K. Dao, and V. T. Nguyen, "Fully-automated person re-identification in multi-camera surveillance system with a robust kernel descriptor and effective shadow removal method," *Image and Vision Computing*, vol. 59, pp. 44 – 62, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885616302189

[4] H. El-Alfy, D. Muramatsu, Y. Teranishi, N. Nishinaga, Y. Makihara, and Y. Yagi, "A visual surveillance system for person re-identification," in *Thirteenth International Conference on Quality Control by Artificial Vision 2017*, vol. 10338. International Society for Optics and Photonics, 2017, p. 103380D.

[5] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1367–1376.

[6] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[7] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[10] 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*, 2015, pp. 91–99.

[11] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2648–2655.

[12] N. Thuy-Binh, T. Duc-Long, L. Thi-Lan, P. Thi Thanh Thuy, and D. Huong-Giang, "An effective implementation of gaussian of gaussian descriptor for person re-identification," in *The 5th NAFOSTED Conference on Information and Computer Science (NICS 2018)*, 2018.

[13] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao, "Set based discriminative ranking for recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 497–510.

[14] Y. Wu, M. Mukunoki, and M. Minoh, "Locality-constrained collaboratively regularized nearest points for multiple-shot person re-identification," in *Proc. of The 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*. Citeseer, 2014.

[15] H. Liu, L. Qin, Z. Cheng, and Q. Huang, "Set-based classification for person re-identification utilizing mutual-information," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 3078–3082.

[16] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 33–40.

[17] 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 (2015)*, 2015, pp. 2197–2206.

[18] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

[19] Q. N. Hong, T.-B. Nguyen, and T.-L. Le, "Enhancing person re-identification based on recurrent feature aggregation network," in *Multimedia Analysis and Pattern Recognition (MAPR), 2018 1st International Conference on*. IEEE, 2018, pp. 1–6.

[20] C. Gao, J. Wang, L. Liu, J.-G. Yu, and N. Sang, "Temporally aligned pooling representation for video-based person re-identification," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4284–4288.

[21] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*. Springer, 2014, pp. 688–703.