Journal of Universal Computer Science, vol. 25, no. 6 (2019), 611-626
submitted: 7/1/18, accepted: 20/5/19, appeared: 28/6/19 © J.UCS

Improving Person Re-identification by Segmentation-Based Detection Bounding Box Filtering

Dominik Pieczyński
(Poznań University of Technology, Piotrowo 3A, 60-965 Poznań, Poland
dominik.pieczynski@put.poznan.pl)

Marek Kraft
(Poznań University of Technology, Piotrowo 3A, 60-965 Poznań, Poland
marek.kraft@put.poznan.pl)

Michał Fularz
(Poznań University of Technology, Piotrowo 3A, 60-965 Poznań, Poland
michal.fularz@put.poznan.pl)

Abstract: In this paper, a method for improving the quality of person re-identification results is presented. The method is based on the assumption, that including segmentation information into re-identification pipeline discards the automated detections that are of poor quality due to occlusions, misplaced regions of interest (ROI), multiple persons found within a single ROI, etc. using a simple segment number, bounding box fill rate and aspect ratio check. Assuming that a joint detector-segmented approach is used, the additional cost associated with the use of the proposed approach is very low.
Key Words: person re-identification, computer vision, deep learning, segmentation
Category: I.2.1, I.2.10, I.4.9, I.5.4

1 Introduction

Person re-identification is one of the most prominent tasks in video surveillance systems. First introduced as a computer vision research problem in the context of human-robot interaction, it was defined as a task to 're-identify a person when it leaves the field of view and re-enters later' [Zajdel et al., 2005]. Since then, it has found its way into video surveillance systems and became popular in the community due to its application and research significance. Pioneering approaches were usually based on colour and texture information. Fusion of information from multiple views followed soon after [Bazzani et al., 2010], along with the approaches that aimed at decreasing the influence of background by performing some kind of segmentation [Farenzena et al., 2010]. Following the recent research trends in computer vision, deep learning based approaches first appeared in [Yi et al., 2014]. Introduction of those new approaches caused a breakthrough change in re-identification accuracy, reaching over 80% rank-1 accuracy on challenging datasets with over 1000 individuals registered across multiple views [Hermans et al., 2017, Li et al., 2017, Zheng et al., 2017].

Figure 1: *Example issues associated with datasets created using automated sample collection – from the left: incomplete person within the region of interest (ROI), misplaced ROI containing a significant portion of background, severe image quality degradation due to blur, occlusion by another person. Images taken from the MARS dataset [Zheng et al., 2016]*

That being said, person re-identification is still a challenging task. Datasets used to train and test the deep learning approaches are based on automatic detection, which might give rise to problems shown in Fig. 1. Although the presented example image pairs are assigned to the same individuals in the dataset, getting correct re-identification using these image pairs would certainly be less likely. On the other hand, a video surveillance system operating under realistic conditions must also cope with these issues.

In this paper, the influence of using segmentation priors on the accuracy of person re-identification is evaluated. The rationale behind the idea is the fact, that in video surveillance systems re-identification is usually preceded by object detection. Since methods for joint object detection and segmentation are available [He et al., 2017], the detection bounding box and its internal segmentation result can be used as an input to a simple rule-based system, that rules out problematic cases so that the re-identification procedure is not executed for potentially invalid image pairs. The influence of the value of key coefficients used for filtering like the detection bounding box aspect ratio and fill rate is investigated in this context. Moreover, two convolutional feature detection backends (one designed for accuracy, and one designed for computational efficiency) performing the re-identification task are compared.

2 State of the art

Current state of the art approaches are evaluated using the Market-1501 dataset [Zheng et al., 2015] containing over 30 thousand images of 1501 individuals and it was considered comprehensive at the time of its introduction. The bounding boxes contained in this dataset are in general of good quality. Although the detections were first performed using an automated approach the detection windows were hand-filtered to some extent. For each detected bounding box to be annotated, a ground truth bounding box containing the pedestrian is drawn by hand. If the overlap between the automated detection and hand annotation is larger than 50%, the bounding box is considered a valid detection. If the ratio is between 50 and 20%, their bounding box is considered a distractor (hard example). Remaining bounding boxes are discarded. The distractors constitute about 8.5% of the dataset. Recent advances enabled a significant improvement of the results on Market-1501 over the baseline performance reported in the paper introducing the dataset (44% in Cumulative Matching Characteristic (CMC) rank-1 accuracy, single query). Using convolutional neural networks as feature extractors in the siamese network setting enabled crossing the 65% accuracy threshold [Varior et al., 2016]. Modifications of the training procedure and the loss function, extensive use of transfer learning and data augmentation soon pushed the results to 80% accuracy and beyond. Usefulness of transfer learning in the context of person re-identification, especially given the Market-1501 dataset's scarcity of data, is demonstrated in [Geng et al., 2016]. An approach using a loss function not directly grounded in image classification and derived from triplet loss is presented in [Hermans et al., 2017]. An approach based on joint training of verification and identification is described in [Zheng et al., 2018]. The approach learns a discriminative embedding and a similarity measurement at the same time. Currently, the highest performance in terms of accuracy (over 93%) is obtained using neural network models incorporating additional knowledge on body parts within an attention mechanism, promoting more holistic silhouette comparison without focusing just on the strongest features [Li et al., 2018][Wang et al., 2018].

However, the rise of data-hungry deep learning solutions naturally led to the introduction of much bigger datasets. The MARS dataset [Zheng et al., 2016] is currently the most comprehensive in terms of number of images – it contains over 1.1 million images of 1261 individuals. Moreover, the images are generated using an automated pedestrian detector, so the detection bounding boxes are imperfect and closer to real-life conditions. The persons within detection windows are often incomplete, or occupy a small fraction of the detection window. Reports of accuracy using this dataset are less common, with [Hermans et al., 2017] being currently in the lead with 80% accuracy as per information on the benchmark's website[1]. The described solution uses the MARS dataset and is based on this top performing method.

[1] `http://www.liangzheng.com.cn/Project/state_of_the_art_mars.html`

3 Proposed approach

The proposed approach is based on two key concepts: the re-identification neural network and prior detection and segmentation operation based on the Mask-RCNN method.

3.1 Re-identification and training

The re-identification neural network model is configured to generate embeddings instead of simply returning similarity measure. This approach is beneficial, since once generated embedding vector can be stored and reused, whereas the similarity measure approaches usually require computationally expensive neural network prediction to be performed for every two images.

The ResNet-50 backend [He et al., 2015] is used as a feature extractor by removing the final classification layer. The network architecture has proven to provide a good balance between complexity and accuracy. Moreover, a lightweight, embedded hardware-friendly MobileNet-V2 neural network was also tested to check the validity of the solution with a significantly simpler (3.4 million parameters vs 25.5 million parameters for ResNet-50) neural network architecture [Sandler et al., 2018]. The use of standard network architectures enables the use of pre-trained models. The feature extraction part is followed by an average pooling layer for final 2048-dimensional embedding computation. As demonstrated in [Lin et al., 2013], average pooling can be successfully applied in place of the fully connected layer, demonstrating better robustness against overfitting with the added benefit of having less trainable parameters.

The model is trained using batch hard triplet loss function. Presented in [Hermans et al., 2017], it is derived from the basic triplet loss introduced in [Weinberger and Saul, 2009]. The basic triplet loss performs optimisation with the aim of transforming the embedding space in a way that makes the data points (e.g. embedding vectors) coming from the same identity (e.g. the same person) closer to each other than points coming from different identities. To perform training, the network is presented with triplets – the anchor image, similar image (belonging to the same identity) and the dissimilar image. While successfully applied to face identification using deep convolutional neural networks [Schroff et al., 2015], triplet loss was outperformed by other approaches in person re-identification. The issue is alleviated by introducing batch hard triplet loss, employing a scheme for random sampling of identities and their corresponding images to form a batch of images. The batch is then mined for the hardest positive and the hardest negative samples, which are subsequently used for the loss function value computation. Doing so, we discard trivial examples, which results in more meaningful updates, and do not rely on hard sample mining within the whole dataset, which speeds up training significantly.

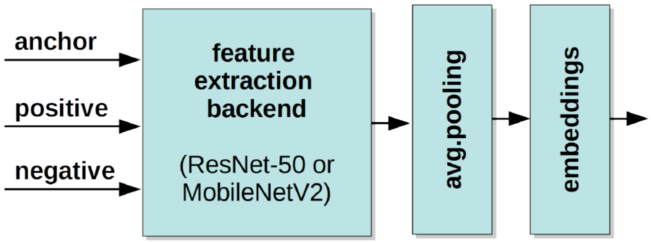

Figure 2: *Block diagram of re-identification neural network architecture. The 'positive' and 'negative' inputs are used only during the training phase*

As demonstrated in [Hermans et al., 2017], using batch hard triplet loss enables significant re-identification accuracy improvement, beating the state of the art approaches at the time of its introduction.

The generalised structure of the network is shown in Fig. 2. The architecture enables straightforward re-identification using computed and stored embeddings. No additional steps or techniques used in re-identification systems such as re-ranking [Zhong et al., 2017], metric learning [Liao et al., 2015] or multiple query [Zheng et al., 2015] were applied.

Adam optimiser [Kingma and Ba, 2014] was used during the training. All hyperparameters were set to default values except for the learning rate of $2e-04$ and weight decay rate of $5e-4$.

A subset of the training part of MARS dataset [Zheng et al., 2016] was used for the training purpose. The training part includes 509 914 pictures of 625 individuals. Only the persons visible by 4 or more cameras were included in the training process. That selection limited the dataset to include 302 172 (59%) pictures of 214 (34%) individuals. The data was randomly split into training and validation sets so that images of 160 persons were used for training and images of 54 persons were used for validation. To benefit from transfer learning to speed up the training and improve the results [Yosinski et al., 2014], we used ResNet-50 and MobileNet-V2 variants that were pre-trained on the ImageNet for initialisation. The network was trained for 300 epochs. The model's weights were automatically saved when there was an improvement in validation accuracy.

3.2 Bounding box contents segmentation with Mask-RCNN

Incorporating segmentation directly into the object of interest detection has several advantages in the context of person re-identification. First, the process of re-identification can be constrained to the object of interest, eliminating the influence of background that can change significantly across multiple views of an object. Second, the coefficients such as ROI fill ratio can be used to rule out

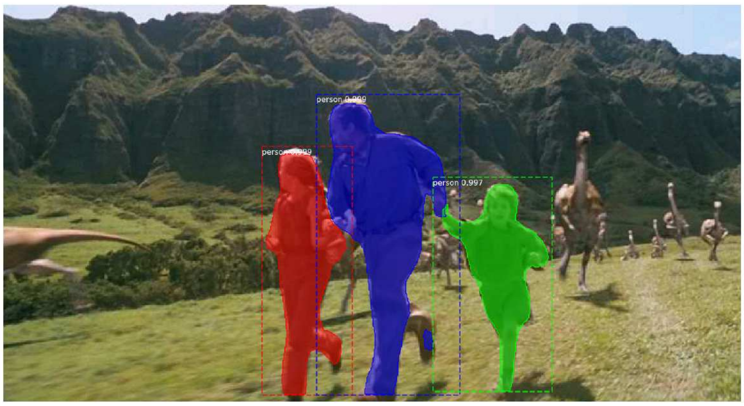

Figure 3: *Mask-RCNN, aside from performing object detection, effectively differentiates between objects and the background, even if the image quality is imperfect*

the imperfect images from processing.

Amongst the novel methods for object detection, Mask-RCNN is especially interesting. It is able to detect multiple entities and perform automated segmentation inside their bounding boxes. This means that using this approach as a detector or an automated method for generating new re-identification dataset extends the standard bounding boxes data with masks that allow testing of a background suppression effectiveness (see Fig. 3).

Mask-RCNN is an extension to Faster-RCNN [Ren et al., 2017] network. It uses the same principles to detect objects, but adds additional, parallel branch that performs segmentation.

The idea behind Faster-RCNN is to detect objects in two stages. The network performs its operations using convolutional feature maps as the input. An important characteristic of such approach is that those maps can be generated by a variety of network models. This makes Faster-RCNN decoupled from network's base and allows using both simpler and more complex models depending on the need and available hardware.

The first stage called Region Proposal Network (RPN) finds a predefined number of rectangular regions that may contain objects. In order to perform this tasks the RPN uses anchors, which are fixed sized bounding boxes that are moved over the image using configurable stride.

As the anchors number, aspect ratios and scales can be adjusted prior to the network's training, this method provides a versatile way to select multiple objects of different shapes.

The second stage of Faster-RCNN, which shares network's weights with the first stage, performs feature extraction for every proposed candidate and pro-

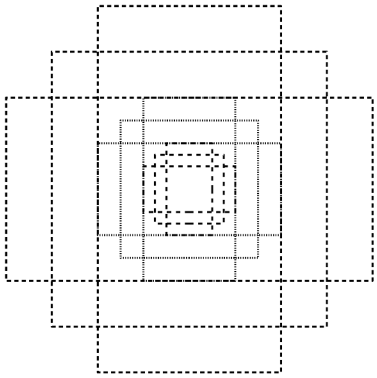

Figure 4: *Faster-RCNN anchors. In this example the network is configured to use 9 anchors (3 bounding boxes in 3 scales denoted by different line styles)*

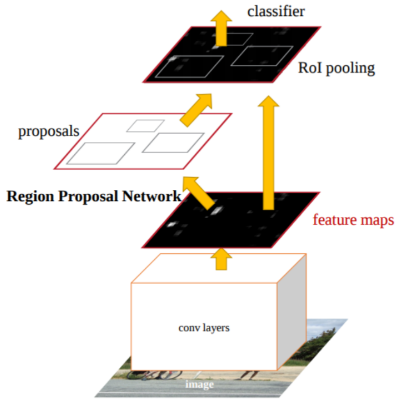

Figure 5: *Faster-RCNN general architecture. Image source: [Ren et al., 2017]*

ceeds with classification as an object or a background and bounding-box regression, which adjusts initial anchor's coordinates.

Mask-RCNN extends the second stage of Faster-RCNN with a parallel, fully convolutional branch that outputs a binary mask for each region of interest. This way the computational demand does not increase dramatically and the network retains Faster-RCNN properties. Most importantly, the *head*, performing bounding-box recognition and mask prediction, is still decoupled from the *backbone*, used for obtaining a feature map.

Mask-RCNN can therefore be used as both the pedestrian detector and the segmenter in re-identification systems. This method can also be extended with

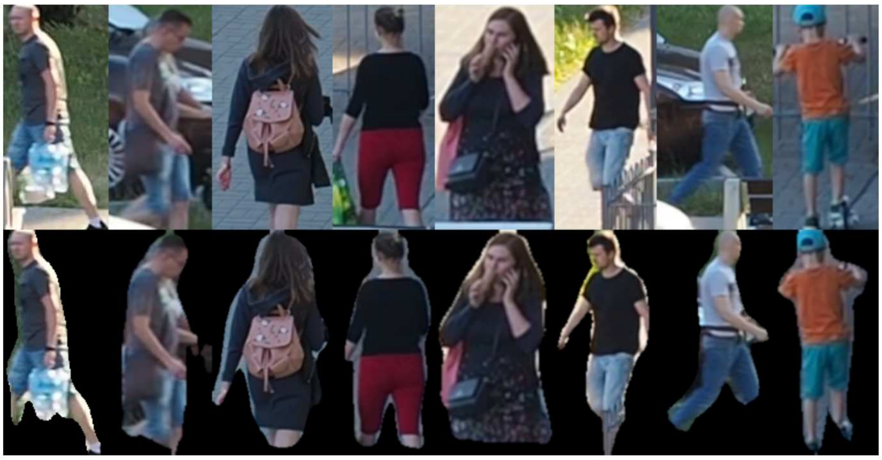

Figure 6: *Example bounding boxes with segmentation results generated fully automatically by Mask-RCNN*

tracking capabilities (retaining person label between video frames). Example results of applying Mask-RCNN in a typical video surveillance setting are shown in Fig 6.

In the video surveillance scenario we are dealing mostly with upright human silhouettes and this is also the most desirable input image to perform re-identification. Since the input image size is scaled to 128×256 ($width \times height$) pixels, the region proposal anchors were set to the sizes of 64, 128 and 256 pixels. The anchor aspect ratios were set to 1.25, 1.6 and 2.0. This gives preference to objects whose height to width ratio is over one for increased domain adaptation. To further accommodate Mask-RCNN to our specific task, additional adjustments were performed. First, the network was trained to detect the 'person' class only, with the above mentioned RPN anchor settings. The COCO (common objects in context) dataset images containing objects of this specific class were used [Lin et al., 2014]. Finally, additional fine-tuning of the last layers of the network (the RPN, the classifiers and the mask heads) was performed. During this last phase the network was trained using the same images, and all but the fine-tuned layer weights were fixed, which resulted in further improvement of accuracy.

3.3 Bounding box filtering

The output of the network (both the detection and the segmentation) can now be used to deal with a range of situations, in which the input data coming from the detection pipeline is not of high enough quality to try and perform

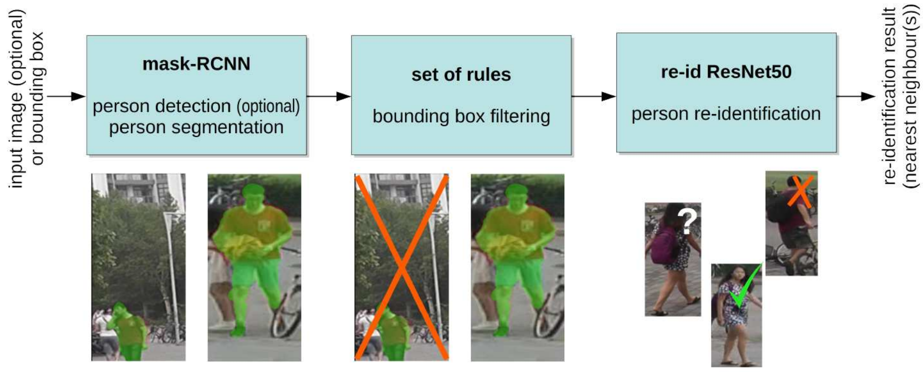

Figure 7: *The flow of data across the functional blocks in the processing pipeline*

re-identificaiton. Such cases include:

1. no object of the 'person' class is detected within the ROI, which may indicate misplaced ROI,
2. more than one object of the 'person' class is detected (with segmentation) within the ROI, which might indicate occlusion by another person,
3. area of segmented foreground (pixels belonging to the 'person' class) within the ROI is too low or too high, which might indicate misplaced ROI or a ROI with a significant portion of background or occlusion by environment elements,
4. the ratio of lengths of the sides of the ROI is too high or too low, which might indicate that no complete person in upright position is present within the ROI.

The influence of all the criteria was evaluated in terms of accuracy increase and the portion of the images discarded. This enables to make an informed decision on how to select the combination of the criteria and their thresholds to achieve the desired accuracy improvement on one hand, and not to discard too large of a portion of the input data on the other hand. The schematic diagram of the flow of data is presented in 7.

4 Evaluation results

The test part of MARS image dataset was used for evaluation [Zheng et al., 2016]. This part of the dataset contains automatically collected data for 636 different individuals. Each individual is observed by up to 6 different cameras, although not all persons are visible in all the cameras. Altogether, the test part of the dataset contains 681 089 images.

For the purpose of this research, a subset of MARS test dataset was chosen. Only test persons visible by 4 or more different cameras were used. This size reduction allows for faster testing of the method. Overall 210 (33%) individuals with 270 475 (53%) images were used. Images were filtered using an implementation of Mask-RCNN [Abdulla, 2017]. The training was performed on a machine with a Titan Xp and Titan V GPU.

The first two detection quality criteria mentioned in the previous section are binary and inform us whether or nor a single person is present within the detection window. The images that didn't fulfil the binary criteria are discarded before proceeding with further evaluation.

To assess the impact of the non-binary criteria (the bounding box fill ratio and the region of interest height/width ratio), the following procedure was used:

– the threshold for fill ratio was increased from 0% to 30% with a 0.5% step increment, recording the accuracy and the number of remaining images for each step,

– the threshold for fill ratio was decreased from 30% to 0% with a 0.5% step increment, recording the accuracy and the number of remaining images for each step,

– the bounding box height/width ratio was increased from 1 to 3 with a 0.05 step increment,

– the bounding box height/width ratio was decreased from 3 to 1 with a 0.05 step increment.

This enables an informed choice of the range of non-binary parameters that results in ignoring problematic cases on one hand, but does not discard too many images from the re-identification process. The results in of the experiments are given in figure 8 and 9. Cumulative matching score was used for evaluation, so $rank{-}n$ means, that the correct person is among the n closest matches, and rank-1 denotes binary accuracy. Evaluation protocol described in [Zheng et al., 2015] and [Zheng et al., 2016] was applied.

The curves are plotted for the ResNet-50 model. To improve clarity, the MobileNet-V2 curves are not shown in the charts, but their shape is roughly similar, with a few percent shift towards lower accuracy, so the key takeout is essentially the same. The range of the fill rate for which the accuracy remains above 0.8 is 15% to 21%. The range of the aspect ratio for which the accuracy remains above 0.8 is 1.9 to 2.4. These ranges were used as valid for non-binary criteria in further tests. Applying the binary and non-binary criteria results in

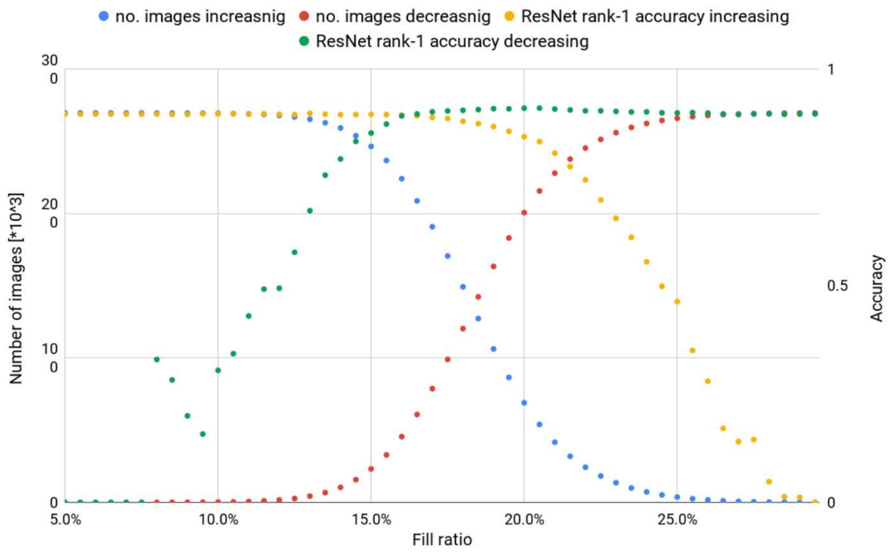

Figure 8: *The chart visualising the fill ratio thresholds influence on the number of images and re-identification accuracy*

discarding roughly 53% of the dataset images. Examples of discarded images are shown in Fig. 10. It should be noted, that while evidently wrong images were removed, a significant portion of fine looking ones were also discarded, since the criteria applied are quite rigorous.

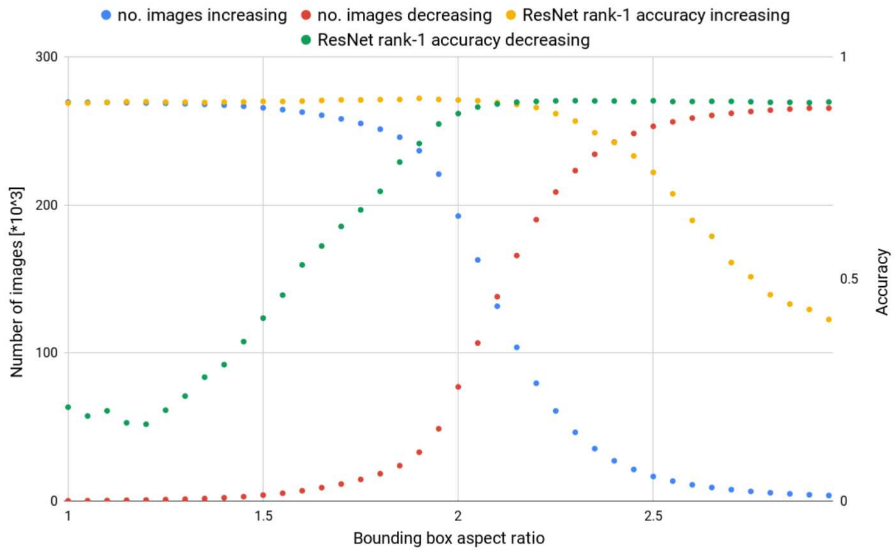

Figure 9: *The chart visualising the bounding box aspect ratio thresholds influence on the number of images and re-identification accuracy*

Figure 10: *Example images automatically removed from test set. Top row detections in the left did not meet aspect ratio criterion, top row detections in the right presented multiple persons and the images in the bottom row, shown as pairs of original image and segmented one, had insufficient fill ratio*

The overall accuracy gains achieved through filtering by applying all three criteria are shown in Table 1. Given the query image, the predictions were performed and the similarity was scored and sorted for all target persons in the database. Rank n accuracy means that the classification was marked successful if the correct person was present in n most similar persons chosen by the network. As MARS dataset requires a random choice of gallery images, we perform 50 passes of accuracy calculation with different selections. In addition to the rank accuracy we also report the standard deviation.

	r-1	**r-5**	**r-10**	**r-20**
ResNet-50 backend				
U	0.896 ± 0.00731	0.938 ± 0.00635	0.950 ± 0.00550	0.961 ± 0.00528
F	0.923 ± 0.00641	0.954 ± 0.00658	0.963 ± 0.00619	0.973 ± 0.00526
MobileNet-V2 backend				
U	0.870 ± 0.00859	0.924 ± 0.00627	0.940 ± 0.00547	0.953 ± 0.00522
F	0.901 ± 0.00811	0.943 ± 0.00661	0.957 ± 0.00586	0.969 ± 0.00463

Table 1: *The tested model with their corresponding rank accuracy with and without input image filtering; **U** indicates the unfiltered dataset, while **F** indicates the dataset filtered with the described rules*

Applying the input image filtering improved the rank-1 accuracy by 2.7 percent points in case of the ResNet-50 backend and by 2.9 percent points in the case of MobileNet-V2 backend. The improvements of higher rank accuracies are significantly lower, as expected. This demonstrates, that the approach that is currently listed as the state of the art level person re-identification accuracy on the MARS dataset benefits from the presented image filtering method. Since the method is generic, one might also expect similar gains in the case of other methods and network architectures. However, the improvements might be not as prominent in the case of methods employing body part attention mechanisms [Li et al., 2018][Wang et al., 2018]. Interestingly, the difference between the two backends used for re-identification is less prominent than in the case of their use for ImageNet classification, in which they achieve 72% and 77.2% accuracy, respectively. This indicates, that an analysis of performance of feature detection backends in deep learning person re-identification might be an interesting research area. This observation is especially valuable in the light of the fact, that video surveillance is increasingly performed using distributed, resource-constrained computational platforms forming smart camera networks

[Shao et al., 2018].

5 Conclusions

A method for improving the accuracy of person re-identification was proposed. The method uses segmentation priors to filter out the problematic images, whose analysis might give rise to errors. The method is based on a variety of simple characteristics, whose computation is possible under the assumption that the joint detection and segmentation approach is used as the prior processing step. The computational cost of computing the aspect ratio and fill ratio is low, yet the improvement in re-identification accuracy is noticeable, even though a state of the art method is used as baseline. Moreover, the method can be used in conjunction with a wide range of existing re-identification approaches. Future work will be focused on observing the interaction between the presented approach and other re-identification performance improvement steps like multiple query or re-ranking. An evaluation of methods based on attention and involving matching of specific silhouette parts is also considered as an interesting direction for research, since the segmentation-based approach presented in this paper is to some extent equivalent. The comparison of deep convolutional feature extractors points reveals, that a more thorough evaluation of a range of other available options might also be valuable.

Acknowledgements

The authors thank Nvidia for hardware donation under Nvidia Academic Hardware Grant.

References

[Abdulla, 2017] Abdulla, W. (2017). Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. `https://github.com/matterport/Mask\_RCNN`.

[Bazzani et al., 2010] Bazzani, L., Cristani, M., Perina, A., Farenzena, M., and Murino, V. (2010). Multiple-shot person re-identification by HPE signature. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1413–1416. IEEE.

[Farenzena et al., 2010] Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE.

[Geng et al., 2016] Geng, M., Wang, Y., Xiang, T., and Tian, Y. (2016). Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*.

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

[Hermans et al., 2017] Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

[Li et al., 2017] Li, W., Zhu, X., and Gong, S. (2017). Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*.

[Li et al., 2018] Li, W., Zhu, X., and Gong, S. (2018). Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294.

[Liao et al., 2015] Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206.

[Lin et al., 2013] Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

[Ren et al., 2017] Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1137–1149.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520. IEEE.

[Schroff et al., 2015] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[Shao et al., 2018] Shao, Z., Cai, J., and Wang, Z. (2018). Smart monitoring cameras driven intelligent processing to big surveillance video data. *IEEE Transactions on Big Data*, 4(1):105–116.

[Varior et al., 2016] Varior, R. R., Haloi, M., and Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer.

[Wang et al., 2018] Wang, H., Fan, Y., Wang, Z., Jiao, L., and Schiele, B. (2018). Parameter-free spatial attention network for person re-identification. *arXiv preprint arXiv:1811.12150*.

[Weinberger and Saul, 2009] Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244.

[Yi et al., 2014] Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE.

[Yosinski et al., 2014] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

[Zajdel et al., 2005] Zajdel, W., Zivkovic, Z., and Krose, B. (2005). Keeping track of humans: Have I seen this person before? In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 2081–2086. IEEE.

[Zheng et al., 2016] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer.

[Zheng et al., 2015] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124.

[Zheng et al., 2017] Zheng, Z., Zheng, L., and Yang, Y. (2017). Pedestrian alignment network for large-scale person re-identification. *arXiv preprint arXiv:1707.00408*.

[Zheng et al., 2018] Zheng, Z., Zheng, L., and Yang, Y. (2018). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13.

[Zhong et al., 2017] Zhong, Z., Zheng, L., Cao, D., and Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327.