https://doi.org/10.2352/ISSN.2470-1173.2019.11.IPAS-268

Improving Person Re-Identification Performance by Customized Dataset and Person Detection

Herman G.J. Groot, Egor Bondarev, and Peter H.N. de With
Eindhoven University of Technology, Dep. Electrical Eng., Video Coding and Architectures Res. Group, Eindhoven, the Netherlands

Abstract

For person re-identification (re-ID), nearly all person re-ID algorithms use public person re-ID datasets, where these datasets all consist of predefined image crops containing a single person. Unfortunately, these image crops are not optimal for video analysis, so that the person detection becomes suboptimal and person re-ID obtains a lower performance score. In this work, several techniques are presented that customize the person images of a popular public person re-ID dataset.

These techniques consist of customization algorithms based on postprocessing the person-detection bounding boxes using the original frames, resulting in several customized datasets to better facilitate person re-identification. We have evaluated five different ways for customization, based on widening the image crops, various aspect ratios and resolutions, and person instance segmentation. We have obtained a significant increase in performance with widened image crops, yielding a convincing performance increase of nearly 3% in the resulting Rank-1 score. Furthermore, when the applied random-cropping process is further optimized to this customization technique, an increase of even more than 4% is obtained. Both performance gains are a strong indication that any future person re-ID system may benefit from customizations based on the original video frames or from specializing the person detector.

Index Terms— Person re-identification, re-ID, person detection, DukeMTMC, DukeMTMC-reID, original *camera output*, image crop widening, fixed aspect ratio, instance segmentation

Introduction

Automated person re-identification is important for numerous interesting applications related to surveillance and human behavior. Person re-identification (re-ID) enables a surveillance system with multiple cameras to automatically determine whether a person that appears in one camera was already previously encountered in another camera. If such a system is deployed with multiple cameras having non-overlapping views, person re-identification becomes more attractive for multiple reasons given below.

For example, when using a smart city-monitoring system having multiple cameras all around the city, person re-ID can then be attractive, when detecting serious undesired behavior or for analyzing crimes captured with video material. More specifically, person re-ID can allow an operator to automatically obtain all previous locations of the suspect and visualize it as a trajectory through the city. This results in important information for event analysis, as there is most likely correlation between the person's trajectory and the trajectories of possible associates, who can also be identified much more effectively.

At present, indoor and outdoor environments are repeatedly captured by surveillance applications to monitor person behavior and crowd flow. Also, the surveillance allows to detect changes in infrastructure, buildings and the state of objects. Hence, there are many areas where person re-ID can be applied and can be beneficial, as person re-ID helps in lowering the burden to track people in multi-camera surveillance systems, especially for longer periods of time.

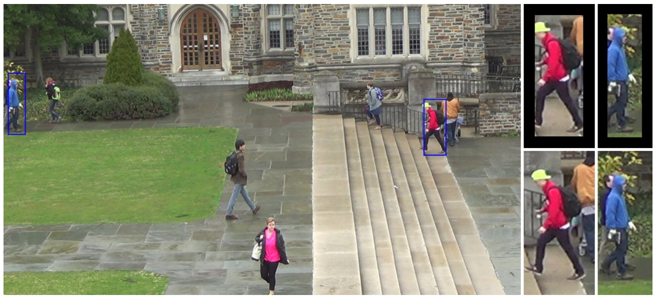


Figure 1. The effect of widening a person image crop in a typical public person re-ID dataset. Two example crops are shown here that are located very tightly around the persons. This happens quite often in any of the most popular public datasets encountered. At the top-right two image crops are shown from DukeMTMC-reID (from camera 2, frame 187077). Their locations in the original frame are indicated by the blue bounding boxes at the left. The black bars at the top-right emphasize the added region of widening, as shown at the bottom-right. When these images (bottom-right) are used instead of the original image crops (top-right), performance increases significantly.

Despite the above advantages, the task of person re-ID is still challenging, due to busy city environments, varying weather conditions and camera viewpoint differences. This explains why only recently acceptable performance scores were reported.

When concentrating on current research, interesting trends become visible. Nowadays, there are many public datasets available, and all adopt the same format. Hence, practically all person re-ID work is based on the same general approach and data setup. On one hand, this is attractive, since all datasets aim to be representative to practical circumstances, capturing persons in a wide variety, with realistic illumination- and pose variations. On the other hand, due to certain assumptions inherently present in all encountered public datasets, person re-ID is still complicated for practical realization.

These assumptions originate from the format of the datasets. To obtain the datasets, firstly, a person detector is applied on all available imagery from all cameras, to obtain bounding boxes of the captured persons. Thus, from this process, only image crops remain. Next, these crops are annotated, i.e. every image crop is manually given a Person ID (PID), to determine which images belong to the same person. The problem of strictly using image crops are multifold and further discussed.

The first problem is that the detection is based on the original camera output, while further processing strictly uses image crops. This blurs the relationship between detection imperfections and the final person re-ID performance score. Second, real-time execution constraints are not considered, whereas this is crucial for many applications. The third problem is that for privacy reasons, an open set of people, where newly encountered people are continuously added is lacking, so that a reliable performance comparison cannot be made.

This paper focuses on the first aspect, researching the influence of the person detection quality on the final person re-ID performance. To this end, we will present several techniques for post-processing the detected person bounding boxes, resulting in customized image crops, which are then used as the input images for the re-ID algorithm. The key to our contribution is that the dataset is customized, such that the dependence on the quality of the person detection becomes visible and can be analyzed.

The structure of this paper is as follows. The next section gives more background information and presents related work from literature. Afterwards, the descriptions of the exact post-processing techniques are provided in a succeeding section. The performances of the presented techniques are presented in the Experimental Results section. Finally, the paper is concluded in the last section.

Related Work

Through the years, several surveys on person re-ID were published [1][2][3]. When Convolutional Neural Networks (CNNs) were introduced, re-ID performance quickly started improving, which hence served as a turning point for person re-ID as well. All recent person re-ID state-of-the-art work employs a CNN and ever few traditional methods with handcrafted features achieve performances near that of a CNN. Nevertheless, many traditional algorithms were published over the years, of which [4]-[7] are promising. Thereafter, approximately at the time of emerging CNN reporting, studies [8][9] were published, where CNNs clearly outperformed conventional algorithms.

A typical CNN produces a likeliness score for every class, where for person re-ID a class is the specific Person ID (PID) in the CNN input image. Consequently, if such an early CNN would be used, the number of output classes (the CNN output vector size) would quickly grow with the number of persons. Even worse, for each new set of people, the CNN would require re-training to maximize performance. To deal with this issue, several solutions can be found in person re-ID literature, roughly divided into verification CNNs and metric-embedding CNNs.

When *verification CNNs* are considered, the CNN takes as input two images of persons and the network then provides a score that describes how likely it is that both images contain the same exact person. Alternatively, when considering *metric-embedding CNNs*, the CNN learns to embed a metric. As such, the CNN takes a single-person image as input and then determines a feature vector, a so-called embedding (of the person properties). Thereby, the distance between the embeddings of images of the same person are supposed to be small (i.e. distance in the embedding space, measured with e.g. the Euclidian distance), while those of different persons are supposed to be relatively high. When compared to a verification CNN, this type requires less computation. That is, since a verification CNN can only focus on two images at a time, it must apply the network for every possible image pair between the current query and the full database. This is more expensive than using a metric-embedding CNN, where at test time, the CNN is applied only to the query image, in order to evaluate its distance to the database embeddings, which are computed in advance. Hence, adopting such a CNN makes a re-ID algorithm more feasible in practice and this explains why all current top-performing re-ID algorithms [10]-[12] are based on metric-embedding networks.

Lastly, the categorization of using global and local features is worth mentioning as well. When using global features, the algorithm obtains a single global feature vector per person image, preferably end-to-end learned by a CNN. For local features, the algorithm obtains one final feature vector per person image that consists of a combination of multiple *partial* feature vectors. In this description, a global feature vector is a vector obtained by looking at the complete person image, while a partial feature vector is obtained by considering some part of the person image, e.g. one of the body parts. For local features, it is also possible that a global feature vector is part of the combination.

Interestingly, research typically shows that combining partial features yields improvements over a single global feature [13]-[16]. Yet, two of the currently available re-ID algorithms [11][12] yield a high performance and are based on using global features. However, the work in [11] represents a fusion of the two categories, as it uses local features, but only during training and not at test time.

Finally, the overall conclusion of studying the related work is that all previously mentioned algorithms have in common that they are based on solely using image crops of all persons. This blurs the relation between the detection imperfections and the final person re-ID performance score. To find this relation, we will present an algorithm based on post-processing of detected bounding boxes, resulting in a customized dataset which leads to higher re-ID performance.

Research Method

This section outlines the processing steps of this study, where the flowchart of these steps is depicted in Figure 2. As illustrated, we have customized the DukeMTMC-reID dataset in several ways. To further clarify the exact differences between our customization techniques, Figure 3 contains a sample of each resulting customized dataset in an enlarged view. To show that typical re-ID algorithms do depend on the quality of the person detection as embedded in the applied datasets, we use the approach that is depicted in Figure 2.

As evident in the diagram of Figure 2, several paths can be taken to determine the re-ID performance score of a path. Each path corresponds to a different dataset customization, and we will refer to each path as a customization variant. Prior to defining each individual path, we first focus on the general aspects applicable to all customization variants. More specifically, in the next subsection Datasets, the necessity of using the DukeMTMC-reID dataset is discussed. Then, in the Preparation Processing subsection, the preparation steps are considered applying to all customization variants. Consequently, these two subsections cover all diagram actions prior to the *Customization Technique* blocks of Figure 2. Thereafter, in the next subsection, the *CNN-based re-ID Algorithm* block of Figure 2 is defined. Next, the optimizations related to the re-ID algorithm hyperparameters are discussed in the corresponding subsection, which concludes the general aspects of all customization variants. Finally, the succeeding subsections discuss all individual customization techniques, six in total.

A. Datasets

Besides the papers, many public person re-ID datasets appeared as well. Nowadays, the most important datasets for person re-ID are Market-1501 [17], CUHK03 [18], and DukeMTMC-reID [19][20]. These are the largest datasets, which is essential for training CNNs. However, there are also many smaller ones (e.g. i-LIDs, VIPeR, PRID, ETHZ, CUHK01, etc.), but they were used only earlier and are not further exploited in this study.

As previously stated, all these datasets have adopted the same data format, based on only image crops without any original frames. However, there is a single exception: DukeMTMC-reID [20]. This

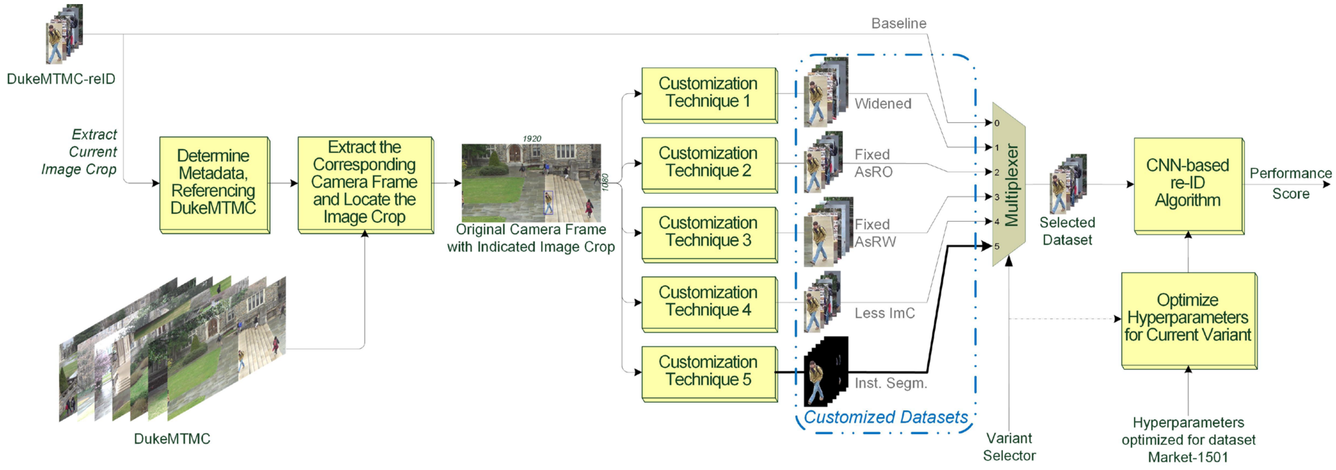

DukeMTMC-reID
Baseline
Extract Current Image Crop
Determine Metadata, Referencing DukeMTMC
Extract the Corresponding Camera Frame and Locate the Image Crop
1920
1080
Original Camera Frame with Indicated Image Crop
DukeMTMC
Customization Technique 1
Customization Technique 2
Customization Technique 3
Customization Technique 4
Customization Technique 5
Widened
Fixed AsRO
Fixed AsRW
Less ImC
Inst. Segm.
Customized Datasets
0
1
2
3
4
5
Multiplexer
Variant Selector
Selected Dataset
CNN-based re-ID Algorithm
Performance Score
Optimize Hyperparameters for Current Variant
Hyperparameters optimized for dataset Market-1501

Figure 2. *Block diagram that shows the processing steps in our study. Prior to training the algorithm, all input images are re-distributed over the customized datasets. The bottom customization output arrow is in bold, because several output variations are used (explained in the text).*

dataset is derived from the person-tracking dataset DukeMTMC [19], which also contains all original camera frames. Hence, this is the reason why we employ this dataset for our experiments to show the importance of the person detector quality.

B. Preparation Processing For All Customizations

The processing blocks in Figure 2 in the region prior to the stack of *Customization Technique* blocks represent the access to the original camera output, which is generally used to create image crops that extract a customized area of the frame.

In more detail, the following is applied for every image crop in DukeMTMC-reID. First, the full frame (original camera output) that contains the current image crop is extracted from the DukeMTMC dataset. Second, the bounding box (bbox) that led to the current image crop is localized and customized, as dictated by the selected *Customization Technique* block. Third, this customized bbox is then used to generate the corresponding customized image crop by extracting the related area of the full frame. Finally, these three steps are repeated for every crop in the DukeMTMC-reID dataset.

In conclusion, the re-ID performance obtained using our customized datasets are virtually fully consistent with the reported results on the original DukeMTMC-reID dataset. This is feasible conclusion because the same original frames are used and the processing is as specified in the original papers. This has been verified by the obtained scores of the various customization methods.

C. Person re-ID Algorithm

To test our hypothesis that the person detector quality has a high impact on the re-ID performance, we have adopted an existing person re-ID algorithm [10]. For the selection leading to this algorithm, several factors were considered. The selected algorithm in [10] is a customized ResNet-50 CNN implementation that uses the triplet loss function to perform deep metric learning and is trained end-to-end. Deep metric learning is an integral part of the whole learning process and is thus not a separate step. The ResNet-50 architecture was customized by discarding the last layer and adding 2 extra fully connected layers. Consequently, even though [10] describes an algorithm that has been succeeded by many alternatives, it proved to be the best choice for adoption in our work because it is in many aspects similar to the actual state-of-the-art and still performing well. Hence, the results presented in this paper will indeed show that the results are comparable with recent person re-ID methods.

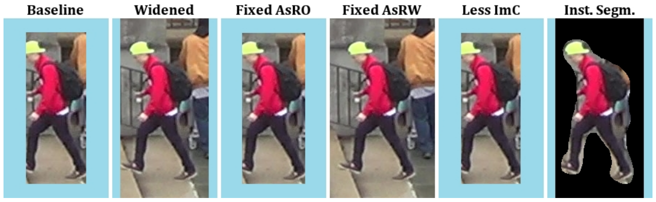


Figure 3. *Sample overview of the investigated dataset customizations related to Figure 2. The blue frames indicate the area of the corresponding full frame (see Figure 1) that is utilized for at least one of the other image crops. The black area of the rightmost crop is part of the CNN input and represents recognized background area.*

D. Hyperparameter Optimization

Since the selected re-ID algorithm was not trained earlier on DukeMTMC-reID, it is necessary to verify that the algorithm is still tuned to this dataset and our newly-generated customization variants (five in total, see Figure 2). For this verification, only our first four variants were considered. Once their performance revealed that tuning the algorithm to each customization variant specifically showed that the hyperparameter values were hardly affected, we omitted the algorithm finetuning for our fifth variant, as it would lead to the same hyperparameter setting. Further details are found in the Experimental Results section, subsection Hyperparameter Settings.

The following algorithm hyperparameters are involved in the algorithm finetuning, as they most likely impact the performance when a different dataset is used [10]: (1) the learning rate, (2) the total number of training iterations, and (3) the number of training iterations after which the learning-rate decay starts.

E. Baseline For All Customization Variants

The original DukeMTMC-reID dataset is used as the baseline for all our customization variants. By comparing the re-ID performance of these variants with the re-ID performance of this baseline, the performance gain can be determined. For convenience, we refer to this baseline as a customization variant in the remainder of this paper.

F. Customization Variant 'Widened'

In many cases, we have found that the image crops are cut so tightly around the actual person (in the full camera frame) that some of their body parts are cut-off. This is not a property of DukeMTMC-reID alone, but appears in practically every public person re-ID dataset that we have studied. Therefore, we investigate here what happens if the image crops contain more of the original image information, i.e. if the crops are widened.

To construct these widened image crops, 20 pixels are added to each of the bbox sides, with limitations when the image borders are too near. This padding of surrounding data ensures that body parts are seldomly cropped off. However, it can still occasionally happen. Typically, this occurs when someone is walking quite fast and is captured when both legs are mostly spread. Widening by 20 pixels is empirically chosen and considered as a balanced choice.

G. Customization Variants With Fixed Aspect Ratio (Fixed AsRO And Fixed AsRW)

Traditionally, CNNs reshape images with variable aspect ratios to a fixed size, thereby changing the original aspect ratio. In case of person re-ID, this affects persons to become less slim than they are (or inverse). Since this corrupts the information on the person shape, it may negatively impact re-ID.

Therefore, the two aspect ratio variants of this subsection ensure that all image crops have a fixed aspect ratio. This is achieved by extending the width of the image crops such that it fits to the height in the desired aspect ratio. As a result, the whole customized dataset has this aspect ratio. Afterwards, the common CNN image-resize operation is still applied to ensure that the CNN is always supplied with images of fixed resolution, as this resizing is part of the person re-ID algorithm. Consequently, the aspect ratios change again, but it is now ensured that every image crop gets the same aspect ratio change from end-to-end.

We have created two variants for the previous (effective) two-step procedure. In the first variant, the image crops from the baseline are extended as described above. We refer to this variant as *Fixed AsRO* to indicate the overall fixed aspect ratio of the resulting customized dataset (Fixed AsRO refers to: Fixed Aspect Ratio Original). For the second variant, the image crops from the customization variant 'widened' are extended with the same method (referred to as *Fixed AsRW:* Fixed Aspect Ratio Widened).

H. Customization Variant With Less Image Compression (Less ImC)

This variant investigates how (higher) image compression may impact performance. We noticed that the image crops from the DukeMTMC-reID dataset reveal slightly different compression artifacts than the original frames (DukeMTMC). That is, the 'blockiness' artifacts from block-based compression in the DukeMTMC-reID image crops do not align with the 'blockiness' artifacts of the original frames. This is most likely caused by the different block-grid location of the crops in the original frames, in combination with applying compression again. Therefore, in this customization variant, the original frame data at the desired image-crop location are directly copied to recreate the image crops of DukeMTMC-reID with the original MPEG compression artifacts.

I. Customization Variant Using Instance Segmentation (Inst. Segm.)

With this customization, we investigate whether instance segmentation (i.e. person segmentation) can help improve re-ID performance, as this forces the network to solely learn the person of interest (PoI, see e.g. the segmented sample in Figure 3). Consequently, instance segmentation ensures that the network cannot coincidentally find some random property in the background that would help the re-ID. This approach as a whole makes it harder for the network to overfit.

However, creating a proper instance segmentation mask for every PoI proved to be complicated. That is, in many image crops multiple persons occur, which results in multiple instance masks per image crop, even with a perfect instance segmentation. Hence, a selection process (see below) is required to localize the PoI for each image crop. Furthermore, the selection process may also be able to resolve some of the instance segmentation errors that occur, but is not always capable to correct those errors.

The exact selection process, as used for this variant, first removes all masks that the instance segmentation classified as a car or an accessory, like a backpack or an umbrella. Unfortunately, sometimes the selection process cannot remove all non-person masks, because the instance mask that best describes the PoI is a misclassified mask. Next, the mask that has the most overlap with the middle area of the image crop is selected. Finally, all accessory masks that are located near the selected middle mask are re-added.

To create the customized dataset, Mask R-CNN [21][22] is applied to every image crop from the widened dataset, where the Mask R-CNN instance was pretrained on the COCO dataset, as a part of [22].

J. Optimizing Crop-Strength Of Random Cropping

As a final step, the algorithm's random cropping phase is finetuned to the best performing dataset customization, which is identified in the next section as Variant Widened. Prior to finetuning, the processing pipeline of the re-ID algorithm is first explained. The first step of the pipeline is resizing of the image crops from the customized dataset to a fixed resolution. The second step is the random cropping phase, where the resized images are cropped to another fixed-, but lower resolution and the remaining area of the image is selected randomly. Finally, the CNN is applied on these randomly-cropped resized images.

Since the CNN input image dimensions are fixed and the algorithm cannot alter the image dimensions of the dataset itself, the first image resizing step effectively determines which part of the image is removed by random cropping. The desired image dimensions form a key parameter that defines the crop-strength of random cropping. In other words, this determines which area of the original image is supplied to the CNN. We refer to this area as the *CNN Window*.

Additionally, this image resizing changes the aspect ratio as well. Therefore, in the experiments we have explored two aspect ratios with four resolutions each. This experiment makes it possible to find the best combination of resolution and aspect ratio for all cropping strategies. The first aspect ratio is based on the original DukeMTMC-reID dataset, while the second aspect ratio is using the default parameters of the re-ID algorithm.

Experimental Results

This section is divided in four subsections. The first subsection (A) describes which hyperparameter setting proved to be optimal for DukeMTMC-reID and our customization variants. The second subsection (B) on Dataset Customizations indicates which customization variant yields the best re-ID performance. The third subsection (C) addresses which crop-strength of random cropping is optimal. Finally, the fourth subsection (D) evaluates instance segmentation.

For every customization variant, several full training iterations are performed to determine the re-ID performance. This repetition is required to obtain stable output results, since every iteration leads to a slightly different performance score, even with constant hyper-parameters.

A. Hyperparameter Settings

In the previous section, it was found that the first four customization variants resulted in the same optimal hyperparameter setting. In the subsection, we explain how we came to this conclusion and describe the underlying experiments.

We first evaluate which value settings of the chosen hyperparameters optimize the selected person re-ID algorithm. This is repeated for every individual customization variant. Table 1 depicts five different settings for the hyperparameter values. Figure 4 visualizes the scores for the five different parameter settings for each of the five shown customization variants. As mentioned above, training is repeated a number of times, therefore the error intervals in Figure 4 indicate the spread of the Rank-1 scores of the repetitions per hyperparameter setting, while the rectangular bars indicate the average values. These averages result from the repeated experiments for every setting to obtain stable outcomes.

To limit the total training time, training on each setting is repeated with a variable number of iterations. For the baseline, each hyperparameter setting was executed 3 times, except for Setting 3, which was executed 14 times. For all other customization variants, it proved sufficient to run Setting 1, 4, and 5 only once, since the performance of these settings compared to the baseline is nearly identical (the rectangular bars of these settings have therefore no error interval for that reason). Finally, Settings 2 and 3 executed 3 and 14 times, respectively.

The gain in Rank-1 performance score of each setting is plotted relative to Setting 3 of that customization variant, because this setting proved to be optimal for all customization variants. Setting 4 comes close in performance to Setting 3, but shows less stable behavior, as it performs sometimes equal and sometimes higher. Furthermore, Setting 3 is in agreement with the hyperparameter values from the selected algorithm in [10], thereby confirming that this default setting is also optimal for DukeMTMC-reID.

B. Performance Of Customizations Variants

Now that the optimal setting of hyperparameters is found, training is repeated 14 times for each variant, see Figure 5 for the obtained results. The figure shows the performance distribution of the repetition iterations of each variant in the form of box plots.

It can be observed that a significant performance increase is obtained by widening the image crops. This confirms that the original person image crops are indeed cut too tightly around the persons. Hence, it is logical to extrapolate that, when a recent state-of-the-art detector would be deployed, the impact on performance would be even higher.

Table 1: The values of the three hyperparameters for every hyperparameter setting, as referenced in Figure 4.

	Learning Rate (LRate)	Number of train iterations	Starting point of LRate decay
Setting 1	$1 \cdot 10^{-4}$	50,000	35,000
Setting 2	$3 \cdot 10^{-4}$	50,000	35,000
Setting 3	$3 \cdot 10^{-4}$	25,000	15,000
Setting 4	$5 \cdot 10^{-4}$	25,000	15,000
Setting 5	$8 \cdot 10^{-4}$	25,000	15,000

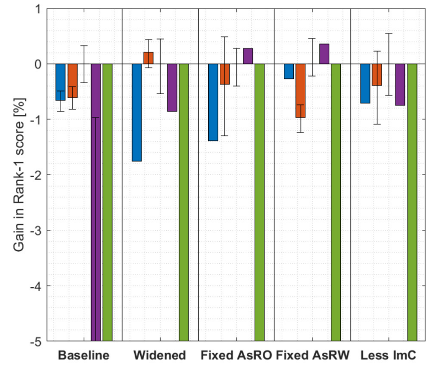


Figure 4. *Gain in re-ID performance for the hyperparameter settings of Table 1, applied for every customization variant. For all rectangular bars of a dataset customization, the gain is shown relative to its Hyperparameter Setting 3 and all negative gain values are clipped at -5% to improve visibility. The error intervals on top of each rectangular bar indicate the 25th and 75th percentile of the repeated iterations.*

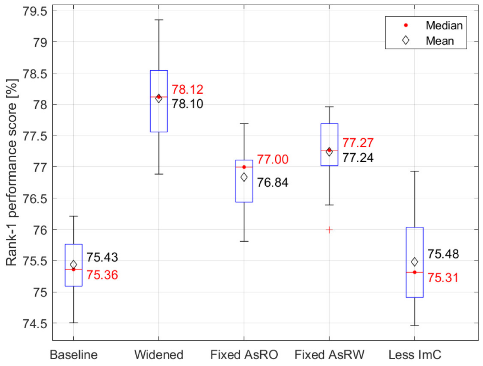


Figure 5. *Rank-1 performance score comparison between the dataset customization types, as described in the Research Method section. The box plots show the distribution of in total 14 repetitions.*

From the other results in Figure 5, it is remarkable that both *fixed aspect ratio* variants do not outperform the *widened* variant. However, they do outperform the *baseline*. This performance increase is likely explained by the effective widening of the image crop due to the aspect ratio adaptation, so that more context pixels are inside the crop. This is in agreement with the *Fixed AsRW* variant performing better than the *Fixed AsRO* variant. However, this variant does also show that the network without any modifications, is not sufficiently able to learn to exploit the size of a person for re-ID. After all, in most cases, the newly added extra pixel columns for these variants contain either background or adjacent persons.

Table 2 – The after-resizing resolutions as used in Figure 6, with similar color coding. The difference with the CNN input dimensions, which is 256 x 128, is removed by random cropping. In the CNN Window column, the effective CNN Window of the input image after random cropping is also indicated. This is relative to the average input dimensions, which is 258 x 124 for DukeMTMC-reID. The two bold table dimensions refer to the entry in the rightmost column.

	After resizing, in [H x W]	CNN Window (w.r.t. input) in [H x W]	Aspect ratio based on
Reso. 1	319 x 199	207 x 80	Non-widening average
Reso. 2	303 x 189	***218 x 84***	
Reso. 3	288 x 180	229 x 88	
Reso. 4	275 x 172	240 x 92	
Reso. 5	298 x 149	222 x 107	Default parameters
Reso. 6	***288 x 144***	229 x 110	
Reso. 7	278 x 139	238 x 114	
Reso. 8	268 x 134	246 x 118	

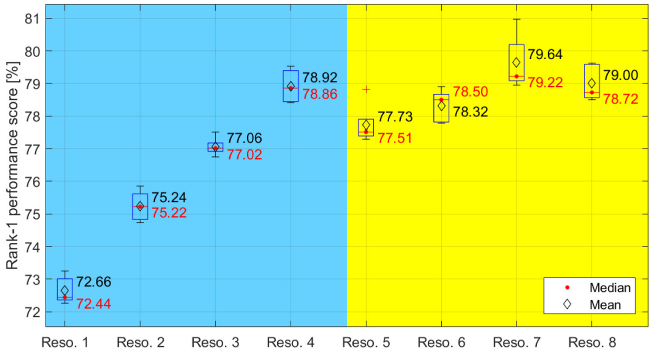


Figure 6. *Influence of the crop-strength of random cropping for the resolutions as shown in Table 2. Bars in similar colored areas use the same aspect ratio.*

However, it is possible that correcting for the person pose with respect to the camera may make the person size more clearly visible.

Finally, with respect to the *Less ImC* variant, it can be concluded that extra image compression has negligible impact on the performance.

C. Crop-Strength Of Random Cropping

Performance can be increased even further when adjusting the crop-strength of the random cropping phase. This, we solely apply random cropping on the *widened* variant, which performed best in Figure 5. As mentioned, the input image is resized during random cropping. In Table 2, the used resizing resolutions and effective CNN Window of the input image are depicted. These resolutions are referenced in Figure 6, which presents the corresponding impact on the random-cropping performance. The CNN Window is chosen relative to the average input image size, which is 258 x 124 pixels for DukeMTMC-reID. If after resizing the dimensions increase, the CNN Window decreases, since the dimensions of the CNN input image (i.e. after resizing and after random cropping) remain constant.

Both choices of the aspect ratio show that random cropping of less pixels improve performance. Furthermore, the default aspect ratio, where the image crops are resized such that their height is double the width, is shown to be most efficient (see the yellow area in Figure 6 and the corresponding values in Table 2).

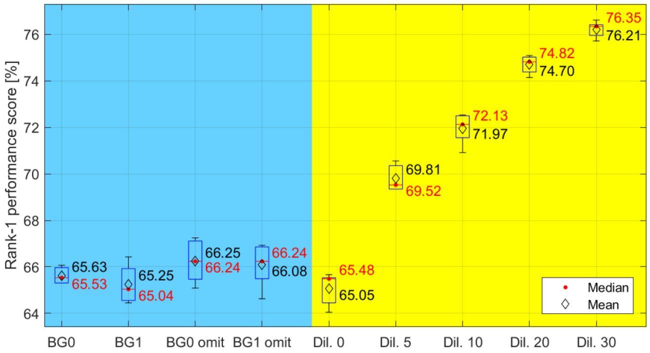


Figure 7. *The instance segmentation results. In the blue area, the influence of either the background value (BG) or the exclusion of faulty segmentations (BG0 omit and BG1 omit) is shown. In the yellow area, the segmentation masks are increasingly dilated when going to the right.*

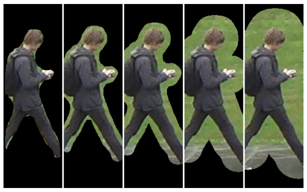


Figure 8. *Impact of dilating the segmentation masks. These image crops correspond to the five dilation levels as indicated in the yellow area of Figure 7, respectively.*

D. Instance Segmentation

Figure 7 presents the results on the instance segmentation dataset. The blue area shows that setting the detected background pixels to either grayscale intensity value 0 or 1 has negligible impact on performance, as all results are approximately equal. Furthermore, a slight increase in performance is evident when we omit the images on which the instance segmentation fails to produce any masks. This indicates that these failure cases are unlikely to be responsible for the observed drop in performance.

The yellow area of Figure 7, representing the final instance segmentation masks are dilated with an increasing number of pixels. The visualization of these dilation levels is shown in Figure 8. The results show that any positive level of dilation increases performance, up to even high dilation levels.

Furthermore, when comparing the scores in Figure 7 with the *widened* variant, it is remarkable that none reach the score of the *widened* dataset. This can either mean that there are too many mistakes present in the instance segmentation such that the network can no longer learn person traits properly, or that the network is utilizing coincidental similarities found in the background that help to match people and thus indicate that the network is overfitting. We consider the first option most likely, since it is difficult to design a selection method that automatically selects the correct mask. This holds particularly when the PoI is occluded by other people.

Discussion

Since DukeMTMC-reID is the only dataset that publishes the original frames, we have reported results on this single public dataset to facilitate comparisons. Although the other datasets have not published the original frames, our reported results do provide a good comparison with related work.

Conclusion

We have customized the public DukeMTMC-reID dataset in five different ways to analyze the dependence of the data input on the operational quality of a person detector that is used for person re-identification (re-ID). These five customizations involve a baseline system with the original tight cropping, a widened cropping, two croppings with initial constant aspect ratio (later modified for fixed resolution by the person re-ID algorithm), and finally a cropping with person instance segmentation. The results indicate that a significant performance increase can be achieved by widening the image crops alone. On the selected person re-ID algorithm, which compares well with related work, we witnessed a convincing increase in performance of nearly 3% Rank-1 score. When optimizing the crop-strength of the random cropping, an increase of even more than 4% is obtained. This is a strong indication that the bounding boxes used to obtain the original person image crops are indeed located too tightly around the actual persons of interest and person detector customizations are useful.

Furthermore, since there is only one public person re-ID dataset that allows access to the original frames (DukeMTMC-reID), is also a strong indication that person detector customizations have not been actively pursued in research. Our study has clearly shown that a good object cropping algorithm can largely affect object re-ID performance.

Acknowledgment

The work in this paper is funded by the European PS-Crimson project, in the framework of the ITEA research program.

References

[1] Bedagkar-Gala A, Shah SK. A survey of approaches and trends in person re-identification. Image and Vision Computing. April 2014; 32(4):270-86.

[2] Zheng L, Yang Y, Hauptmann AG. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984. October 2016.

[3] Karanam S, Gou M, Wu Z, Rates-Borras A, Camps O, Radke RJ. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. IEEE Transactions on Pattern Analysis & Machine Intelligence. (1):1-1. February 2018
NB: this is the 2018 revision, for all versions, see arXiv.

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

[5] Ma B, Su Y, Jurie F. Bicov: a novel image representation for person re-identification and face verification. In British Machine Vision Conference, September 2012.

[6] Zhao R, Ouyang W, Wang X. Learning mid-level filters for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2014, (pp. 144-151).

[7] Yoon S, Khan FM, Bremond F. Efficient Video Summarization Using Principal Person Appearance for Video-Based Person Re-Identification. In The British Machine Vision Conference (BMVC), September 2017.

[8] Ahmed E, Jones M, Marks TK. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, (pp. 3908-3916).

[9] Xiao T, Li H, Ouyang W, Wang X. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, (pp. 1249-1258).

[10] Hermans A, Beyer L, Leibe B. In Defense of the Triplet Loss for Person Re-Identification. arXiv preprint arXiv:1703.07737. March 2017.

[11] Zhang X, Luo H, Fan X, Xiang W, Sun Y, Xiao Q, Jiang W, Zhang C, Sun J. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. arXiv preprint arXiv:1711.08184. November 2017.

[12] Almazan J, Gajic B, Murray N, Larlus D. Re-ID done right: towards good practices for person re-identification. arXiv preprint arXiv:1801.05339. January 2018.

[13] Zheng L, Huang Y, Lu H, Yang Y. Pose Invariant Embedding for Deep Person Re-identification. arXiv preprint arXiv:1701.07732. January 2017.

[14] Li D, Chen X, Zhang Z, Huang K. Learning Deep Context-Aware Features over Body and Latent Parts for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, (pp. 384-393).

[15] Su C, Li J, Zhang S, Xing J, Gao W, Tian Q. Pose-driven Deep Convolutional Model for Person Re-identification. In IEEE International Conference on Computer Vision (ICCV), October 2017, (pp. 3980-3989).

[16] Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Wang X, Tang X. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, (pp. 1077-1085).

[17] Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, December 2015, (pp. 1116-1124).

[18] Li W, Zhao R, Xiao T, Wang X. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014, (pp. 152-159).

[19] Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In European Conference on Computer Vision (ECCV), October 2016, (pp. 17-35).

[20] Zheng Z, Zheng L, Yang Y. Unlabeled Samples Generated by Gan Improve the Person Re-identification Baseline in Vitro. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 2017 (pp. 3754-3762).

[21] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 2017 (pp. 2980-2988).

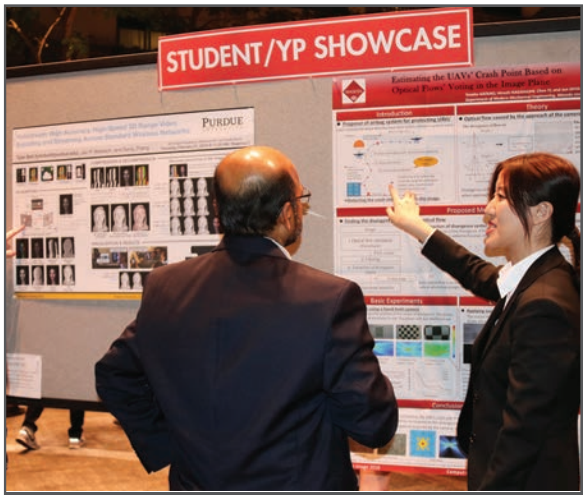
[22] Abdulla W. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. In GitHub repository, available from: https://github.com/matterport/Mask_RCNN, since October 2017, last accessed on 2018 July 17.

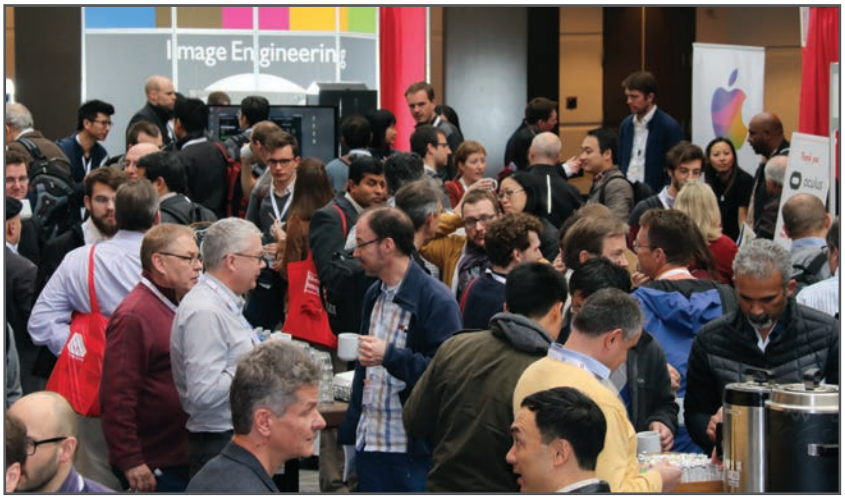
Author Biographies

Herman Groot is a PhD at the Electrical Engineering faculty of Eindhoven University of Technology (TU/e, the Netherlands). His PhD study currently focuses on person re-identification, but halfway through his PhD, the focus will shift more towards robotics, since – ultimately – he strives to be involved in future space-exploration missions. To this end, he would eagerly want to broaden his image processing skills in order to become an expert in space-related image processing techniques. Fittingly, he finalized several MSc elective courses at the Aerospace Engineering faculty of Delft University of Technology (TU Delft, the Netherlands) and did his MSc internship at the Netherlands Aerospace Centre in Amsterdam (NLR, the Netherlands).

Egor Bondarev obtained his PhD degree in the Computer Science Department at TU/e, in research on performance predictions of real-time component-based systems on multiprocessor architectures. He is an Assistant Professor at the Video Coding and Architectures group, TU/e, focusing on sensor fusion, smart surveillance and 3D reconstruction. He has written and co-authored over 50 publications on real-time computer vision and image/3D processing algorithms. He is involved in large international surveillance projects like APPS and PS-CRIMSON.

Peter H.N. de With is Full Professor of the Video Coding and Architectures group in the Department of Electrical Engineering at Eindhoven University of Technology. He worked at various companies and was active as senior system architect, VP video technology, and business consultant. He is an IEEE Fellow, has (co-)authored over 400 papers on video coding, analysis, architectures, and 3D processing and has received multiple papers awards. He is a program committee member of the IEEE CES and ICIP and holds some 30 patents.





AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!