CNN-based thermal infrared person detection by domain adaptation

Christian Herrmann[a,b], Miriam Ruf[a], and Jürgen Beyerer[a,b]

[a]Fraunhofer IOSB, Karlsruhe, Germany
[b]Vision and Fusion Lab, Karlsruhe Institute of Technology KIT, Karlsruhe, Germany

ABSTRACT

Imaging sensors capturing the surroundings of an autonomous vehicle are vital for its understanding of the environment. While thermal infrared cameras promise improved bad weather and nighttime robustness compared with standard RGB-cameras, detecting objects, such as persons, in thermal infrared imagery is a tough problem because image resolution and quality is typically far lower, especially for low-cost sensors. Currently, deep learning based object detection frameworks offer an impressive performance on high-quality images. However, applying them to low-quality data in a different spectral range causes significant performance drops. This work proposes a strategy to make use of elaborate CNN-based object detector frameworks which are pre-trained on visual RGB images. Two key steps are undertaken: First, an appropriate preprocessing strategy for the IR data is suggested which transforms the IR data as close as possible to the RGB domain. This allows pre-trained RGB features to be effective on the novel domain. Second, the remaining domain gap is addressed by fine-tuning the pre-trained CNN on a limited set of thermal IR data. Different IR preprocessing options are explored, each addressing a different aspect of the domain gap between thermal IR and RGB data. Examples include dynamic range, blur or contrast. Because no preprocessing can cover all aspects alone, providing preprocessing combinations to the CNN allows addressing more than one aspect at once and further improves the results. Experiments indicate significant person detection improvements on the public KAIST dataset with the optimized preprocessing strategy.

Keywords: thermal infrared, LWIR, object detection, CNN, domain gap

1. INTRODUCTION

Thermal infrared cameras promise a higher robustness to certain challenges, such as nighttime or adverse weather conditions, compared with standard RGB cameras. This allows autonomous vehicles to sense their environments more reliably under changing conditions. A current drawback of thermal sensors compared with their visual counterparts is a lower spatial resolution and, in addition, worse image quality. Mainly noise is responsible for the low quality which is especially true for low-cost sensors which are the preferred option for most autonomous systems.

In comparison to the surveillance and security domain where thermal cameras are also popular, there are two key differences for autonomous vehicle applications. First, camera price and size is more critical which leads to the choice of small low-cost sensors which provide lower image quality. Second, the addressed environment is an urban scenario where the background is cluttered and has also inconsistent temperatures. Both effects lead to a significantly lower contrast of persons to background compared with surveillance applications which are represented by datasets such as the OTCBVS OSU ones[1,2]. On standard RGB imagery, well known deep learning based object detectors, such as Faster R-CNN[3] or SSD[4] and their derivatives[5], are current state-of-the-art solutions and achieve impressive results. They are often easily available as ready to use pre-trained models. The obvious idea of downloading and running them on thermal infrared data, however, leads to an underwhelming performance in the IR domain.

When trying to improve the results, a key issue of thermal imagery in comparison to the popular RGB domain is the low amount of available data for training. In this work, we want to address this by proposing strategies to employ CNN-based detectors which are pretrained on RGB data.

Two strategies to reduce the domain gap are explored. First, the thermal image data is shifted towards the RGB domain by appropriate preprocessing strategies. Second, the well-known strategy of network parameter fine-tuning with limited

Further author information:
C. Herrmann: christian.herrmann@iosb.fraunhofer.de
M. Ruf: miriam.ruf@iosb.fraunhofer.de
J. Beyerer: juergen.beyerer@iosb.fraunhofer.de

training data from the target domain is combined with the preprocessing strategy. Together, both strategies address the domain gap from both ends: data side and detector side, leading to a convergence between data and detector to improve detection results.

2. RELATED WORK

The general task of object detection is currently dominated by deep learning approaches. Methods, such as YOLO[6] , Faster R-CNN[3] , SSD[4] or RefineDet[5] , are highly popular choices because of their efficiency. Regarding person detection, a key aspect and difference is often the relatively small size of the object with regard to the whole image. For example, current state-of-the-art person detection datasets show an approximate person to image height ratio of about 1:8. This holds for the Caltech[7] as well as the KAIST[8] dataset.

Regarding approaches for person detection in thermal infrared images, a lot of traditional thermal infrared approaches have significant constraints, such as a fixed camera and simultaneously moving persons[2,9–12] , which is required for the often applied background subtraction strategies. Methods based on non-temporal features, such as the gradient, were rare for the thermal infrared domain[13–15] before deep learning solutions started to spread[16] .

The KAIST dataset can be considered the current state-of-the-art dataset dataset for thermal person detection under challenging conditions. It offers significant difficulties caused by the low-cost thermal sensor which results in low contrast, low resolution and significant image noise. Its positive aspect is the parallel availability of RGB and thermal images which current record holders[16] exploit by using both spectral domains in parallel for person detection. In contrast, this work will only use the thermal images from the KAIST dataset to show the benefit of the proposed methods in thermal-only scenarios. This allows a wider application of the proposed methods in simpler sensor setups. On the other side, approaches fusing visual and infrared clues[2,11,16] often achieve better results than pure IR approaches but have consequently a limited applicability.

Besides the straight forward application of standard detection frameworks, the IR domain allows differing methods if high-contrast images are available. Creating candidate proposals by efficient strategies such as MSER[14] in the beginning, and afterwards performing a highly accurate classification of the proposals by Convolutional Neural Networks (CNNs) is an option[17] . These approaches tend to fail, however, for the KAIST low-contrast images. This holds especially at daytime when temperature differences between bodies and the background environment can be low.

3. METHOD

The key aspect which is examined in this work is the reuse of a CNN-based person detector, which is pretrained on RGB data, for thermal infrared person detection. As exemplary detector, the SSD[4] with VGG16[18] as base network is chosen. Because we want to benefit from pretrained RGB CNNs, no network training from scratch will be performed in this work. Simply applying an RGB detector to thermal images provides poor results because of the domain gap between the visible (model) and thermal (data) spectral range. To reduce this gap between input data and pretrained detector model, one can either address the data or the model. In the first step, we suggest to transform the thermal image data in a way that it is shifted closer to the visible domain. Then in the second step, the detector model is adjusted to further reduce the remaining gap.

3.1 Preprocessing

Assuming a thermal image $\boldsymbol{x} \in [0,1]^{wh}$ with width w and height h, we are looking for operations $f : \boldsymbol{x}_p = f(\boldsymbol{x})$ which reduce the gap to the RGB domain. The processed images $\boldsymbol{x}_p$, which are expected to look more similar to RGB images, are then used as input image for the RGB-pretrained person detector. Therefor, the single channel thermal image is broadcasted into all three input channels of the detector. The following four preprocessing operations are suggested:

- Inversion. The most obvious cues that this is a meaningful operation are the sky and the persons themselves. RGB images or grayscale versions thereof show a bright sky with people being usually darker than the background. In thermal infrared images this is usually inverted. Thus, an inversion transforms the thermal images closer to the RGB domain:

$$f_i : \boldsymbol{x}_p = \mathbf{1} - \boldsymbol{x} \; . \tag{1}$$

Figure 1. Different thermal infrared images from the KAIST dataset (rows). Different preprocessing is shown from left to right: none, stretching, equalization, inversion and combination of stretching and inversion.

- Blur. Because low-quality thermal images contain more noise than typical RGB images, high frequency noise can be removed by low-pass filtering the images. For this purpose a Gaussian kernel g is applied with G representing the according Toeplitz matrix:

$$f_b : \boldsymbol{x}_p = G \cdot \boldsymbol{x} \; . \tag{2}$$

- Histogram stretching. The distribution of pixel values in thermal images is often different than for RGB images. Especially for the common 12-bit sensors which capture a wide spectral range. Single extreme hot or cold spots which result in extremely bright or dark spots in the image appear more often than similar effects in RGB images. To allow a comparable contrast, a relaxed histogram stretching is applied. Instead of the strict stretching

$$f_s : \boldsymbol{x}_p = \frac{\boldsymbol{x} - \min \boldsymbol{x}}{\max \boldsymbol{x} - \min \boldsymbol{x}} \; , \tag{3}$$

  $\min \boldsymbol{x}$ is replaced with the threshold for the $\beta > 0$ percentile and $\max \boldsymbol{x}$ accordingly with the $1 - \beta$ percentile threshold. This handles the mentioned outlier spots in the image values and allows for a more stable normalization. We found $\beta = 0.003$ to be a good choice.

- Histogram equalization. This follows mainly the same motivation as the histogram stretching and addresses the same issue. In comparison, it usually results in a higher contrast at the cost of over- or under-saturated image areas.

Figure 1 shows some preprocessed samples. It is easy to see that especially the inversion generates results which are more familiar to the human eye. This indicates being closer to the regular RGB domain which the human vision is trained for. Histogram modifications in shape of stretching and equalization both increase the contrast significantly, however, it is difficult to judge visually which one allows better person detection. Stretching shows less contrast but equalization generates a lot of over- and under-saturated image areas.

3.2 Fine-tuning

A key motivation to reduce the domain gap is the lack of large-scale training data for the thermal domain. However, it is usually feasible and cheap enough to gather a small set of training data in the thermal domain, if not already available. This limited amount of data is usually insufficient to train a detector from scratch. However, it can be used to fine-tune a pretrained detector after preprocessing in order to further reduce the domain gap between images and detector.

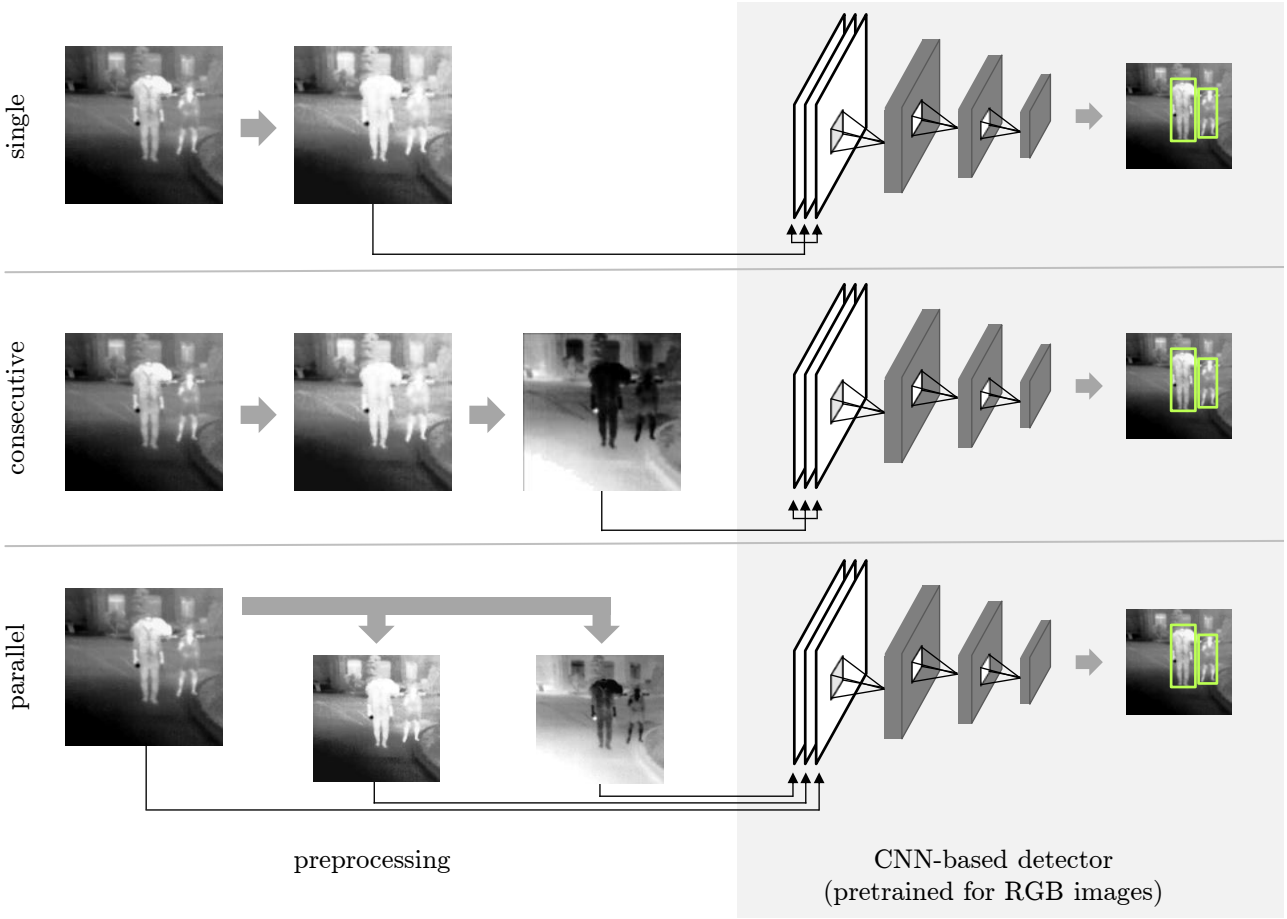

Figure 2. Comparison of preprocessing combination methods. Top: no combination, straight forward application of a single preprocessing method and propagation of single channel thermal image into all three input channels of the RGB pretrained detector. Center: consecutive application of more than one preprocessing method. Bottom: up to three parallel preprocessing methods and distributing the preprocessed images across the three detector input channels.

Generally speaking: preprocessing shifts the thermal image data towards the RGB domain whereas fine-tuning shifts the RGB detector towards the thermal domain. Both effects accumulate as they target a different part of the processing chain and they finally minimize the domain gap between input images and the detector.

3.3 Combination

An obvious next step is the combination of several beneficial preprocessings. Two options are considered: consecutive and parallel data preprocessing as illustrated in figure 2.

3.3.1 Consecutive Preprocessing

The first option is the application of multiple beneficial preprocessing options in sequence to reduce the domain gap as far as possible by gradually minimizing the difference. This assumes that the combined preprocessings are complementary so that their benefit will accumulate. The resulting preprocessed single-channel thermal image is then copied to all three input channels of the detector as before. This allows to benefit from the different preprocessing options at once.

3.3.2 Parallel Preprocessing

As it is unclear if a single consecutive application of preprocessing options is sufficient to minimize the domain gap between thermal and RGB imagery, the second promising option is to offer differently preprocessed images to the detector. Because pretrained detectors for RGB data offer three input channels, three different preprocessings can be inserted in parallel.

Table 1. Detection results on KAIST reasonable all test set for denoted image data. Note that lower values indicate better results.

category	preprocessing	finetuning	log-average miss rate
baseline	none	-	70.47
	none	-	95.51
basic preprocessing	invert	-	78.78
	histogram stretching	-	95.32
	histogram equalization	-	94.12
	blur	-	96.21
	invert + stretching	-	77.64
	invert + stretching + blur	-	78.14
	invert + equalization	-	**76.60**
	invert + equalization + blur	-	76.98
domain adaptation	invert	✓	73.48
	none	✓	70.45
	invert + stretching	✓	**70.25**
parallel combination	none | invert | equalization	✓	**69.81**

Theoretically, this option provides more information to the detector if the preprocessing operations generate opposing information which is lost by consecutive application. Further note that this option is only promising in combination with fine-tuning because intensity relations between the preprocessed channels can be significantly different than relations between RGB channels. Of course, this can be combined with consecutive preprocessing by applying more than one preprocessing operation on one channel.

4. EXPERIMENTS

The experiments are performed with the public KAIST[8] dataset. It offers thermal infrared images along with RGB data of road scenes. Altogether it includes around 95K image pairs with about 86K person annotations. There are officially defined training and test sets which consist each approximately of half the data. Testing is performed with the reasonable all setting which is based on every 20th frame resulting in 2,252 test images. Evaluation is performed with the official scripts which allows comparison to other published results. Note, however, that thermal only results are rarely published for KAIST. For fine-tuning the models, every 20th frame of the KAIST training set is used.

Table 1 indicates the findings starting with baseline results by simply applying the pretrained SSD300 on either the RGB or IR images. Note the big performance drop when applying the detector to IR data. Preprocessing the images boosts the performance significantly and inversion is the single preprocessing which offers the largest benefit. By this strategy, a major part of the performance difference between RGB and thermal is recaptured. The consecutive combination of preprocessing strategies is proven beneficial. Notably, the effect of histogram stretching and equalization appears to be bigger in the combination with inversion than alone. Blurring decreased the results consistently on the KAIST dataset. However, we made the opposite observation on some private datasets with a higher noise level than KAIST. We suspect that the alignment process of the IR data with the RGB data in the KAIST dataset which also includes an upsampling of the raw IR data already involves some steps with sufficient low-pass character. Thus, a further low-pass filtering is proven unnecessary.

Finetuning on the IR test set of KAIST is performed for the most relevant cases. It improves the performance of the detector to the range of direct application of the RGB-detector to RGB data. Note that the results differ depending on the preprocessing with similar findings compared to the non-finetuning case. At first glance this might surprise because, in theory, for the inversion, the network is capable to learn the inversion itself. However, the network is pretrained on RGB data and an inversion would have to be compensated ideally in the filter parameters of the first convolutional layer in the network. This would require a lot of training data and iterations to do this robustly during back-propagation. Due to the limited IR data, this process cannot be fully completed, leading to the different results despite finetuning. The benefit of the histogram processing methods is again notable but smaller than for the inversion.

Providing orthogonal preprocessings in parallel to the detector via separate image channels provides the best results. This can easily be explained because the net can choose the best representation dynamically during training. Note that

there is no conflict with these results being better than the pure RGB ones. Besides the finetuning effect, the KAIST dataset contains some night scenes where persons can be detected better in thermal IR imagery. This leads to a different achievable performance level for RGB and IR data.

5. CONCLUSION

Altogether, the results indicate that the domain differences between RGB and thermal IR data can be mitigated well by the small amount of proposed steps based on RGB-pretrained detector networks. Appropriate data preprocessing, especially inversion, combined with detector finetuning leads to competitive results, even without a large set of domain specific training data. First, making the images look more similar to grayscale converted RGB images allowed feasible results with the pretrained SSD300 detector. Second, combining different preprocessings and fine-tuning pushed the IR results to the level of RGB imagery. Altogether, the proposed strategies showed how to make use of an RGB pretrained detector for thermal infrared person detection.

REFERENCES

[1] Davis, J. W. and Keck, M. A., "A two-stage approach to person detection in thermal imagery," in [*Workshop on Application of Computer Vision*], (2005).

[2] Davis, J. W. and Sharma, V., "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding* **106** (2007).

[3] Ren, S., He, K., Girshick, R., and Sun, J., "Faster R-CNN: Towards real-time object detection with region proposal networks," in [*Advances in Neural Information Processing Systems*], 91–99 (2015).

[4] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "SSD: Single shot multibox detector," in [*European Conference on Computer Vision*], 21–37, Springer (2016).

[5] Zhang, S., Wen, L., Bian, X., Lei, Z., and Li, S. Z., "Single-Shot Refinement Neural Network for Object Detection," *arXiv preprint arXiv:1711.06897* (2017).

[6] Redmon, J. and Farhadi, A., "YOLO9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242* (2016).

[7] Dollár, P., Wojek, C., Schiele, B., and Perona, P., "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34** (2012).

[8] Hwang, S., Park, J., Kim, N., Choi, Y., and Kweon, I. S., "Multispectral Pedestrian Detection: Benchmark Dataset and Baselines," in [*Conference on Computer Vision and Pattern Recognition*], (2015).

[9] Leykin, A., Ran, Y., and Hammoud, R., "Thermal-Visible Video Fusion for Moving Target Tracking and Pedestrian Classification," in [*Conference on Computer Vision and Pattern Recognition*], (2007).

[10] Dai, C., Zheng, Y., and Li, X., "Layered Representation for Pedestrian Detection and Tracking in Infrared Imagery," in [*Conference on Computer Vision and Pattern Recognition Workshops*], (2005).

[11] Elguebaly, T. and Bouguila, N., "A Nonparametric Bayesian Approach for Enhanced Pedestrian Detection and Foreground Segmentation," in [*Conference on Computer Vision and Pattern Recognition Workshops*], (2011).

[12] Chen, B., Wang, W., and Qin, Q., "Robust multi-stage approach for the detection of moving target from infrared imagery," *SPIE Optical Engineering* **51** (June 2012).

[13] Li, W., Zheng, D., Zhao, T., and Yang, M., "An Effective Approach to Pedestrian Detection in Thermal Imagery," in [*International Conference on Natural Computation*], (2012).

[14] Teutsch, M., Müller, T., Huber, M., and Beyerer, J., "Low resolution person detection with a moving thermal infrared camera by hot spot classification," in [*Conference on Computer Vision and Pattern Recognition Workshops*], 209–216, IEEE (2014).

[15] Zhang, X. and Gao, Y., "Face recognition across pose: A review," *Pattern Recognition* **42**(11), 2876–2896 (2009).

[16] König, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., and Teutsch, M., "Fully Convolutional Region Proposal Networks for Multispectral Person Detection," in [*Conference on Computer Vision and Pattern Recognition Workshops*], 243–250, IEEE (2017).

[17] Herrmann, C., Müller, T., Willersinn, D., and Beyerer, J., "Real-Time Person Detection in Low-Resolution Thermal Infrared Imagery with MSER and CNNs," in [*Proc. SPIE 9987, Electro-Optical and Infrared Systems: Technology and Applications*], International Society for Optics and Photonics (2016).

[18] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," in [*International Conference on Learning Representations*], (2015).