

Fusion of Depth and Thermal Imaging for People Detection

Weronika Gutfeter and Andrzej Pacut

Biometric and Machine Intelligence Laboratory, Research and Academic Computer Network (NASK)

<https://doi.org/10.26636/jtit.2021.155521>

Abstract—The methodology presented in this paper covers the topic of automatic detection of humans based on two types of images that do not rely on the visible light spectrum, namely on thermal and depth images. Various scenarios are considered with the use of deep neural networks being extensions of Faster R-CNN models. Apart from detecting people, independently, with the use of depth and thermal images, we proposed two data fusion methods. The first approach is the early fusion method with a 2-channel compound input. As it turned out, its performance surpassed that of all other methods tested. However, this approach requires that the model be trained on a dataset containing both types of spatially and temporally synchronized imaging sources. If such a training environment cannot be setup or if the specified dataset is not sufficiently large, we recommend the late fusion scenario, i.e. the other approach explored in this paper. Late fusion models can be trained with single-source data. We introduce the dual-NMS method for fusing the depth and thermal imaging approaches, as its results are better than those achieved by the common NMS.

Keywords—*depth imaging, person detection, sensors fusion, thermal imaging*.

1. Introduction

The primary goal of this work is to explore the feasibility of detecting human silhouettes in non-visible light spectrum images, without accessing RGB images for reference. Our experiments focus on thermal and depth images showing people in indoor and outdoor environments, i.e. images that are similar to surveillance footage.

The use of alternative imaging sources in computer vision-related tasks is important for numerous reasons. One of those reasons is that they extend the spectrum of features that can be recognized. In this case, recognition is based on the temperature of objects (thermal imaging) and on their geometrical features (depth imaging). Temperature measurements may be critical in the context of the recent pandemic and the demand for wide-scale systems capable of monitoring health parameters. Privacy is another essential aspect that needs to be taken into consideration. At certain location, the use of standard RGB cameras may be prohibited to protect the privacy of data subjects. In such circumstances, surveillance systems relying on alternative

vision cameras may prove to be the best solution available. When working with alternative vision systems, the fact that fewer resources are available than in the case of RGB-image based architectures (data needed to train the algorithms or evaluation benchmarks) is the key challenge.

Consumer-grade non-RGB detectors are usually characterized by lower resolution levels, and the images have poorer quality than their RGB counterparts. That may affect the precision of detection. The size of thermal images used in this paper is 160×120 pixels and the average size of the detection boxes framing human silhouettes is approximately 36×56 pixels. Still, detection performance of thermal images surpasses that of higher resolution depth images (see Tab. 1) with the resolution of the latter equaling 1280×720 pixels. This is probably caused by a higher level of noise in depth images which hampers their depth accuracy. Therefore, a fusion of different image sources may lead to the improvement in results. Specific methods relied upon for merging thermal and depth imaging will be discussed in the second part of the paper.

The investigation of the ability to determine the correct fusion methodology required that a dataset be identified containing images of both types, with the pairs of images being spatially and temporally aligned. These requirements are met by the IPHD dataset which was complied for the Identity Preserving Human Detection Challenge [1] organized in 2020. The IPHD dataset was built using frames extracted from two synchronized image streams: a thermal one and the other containing depth-related information. The dataset is used to evaluate the detection methods proposed in this work and to train the models. We do not use any auxiliary data for estimating models' weights. However, we employ transfer learning techniques from models that were pre-trained using the Common Objects in Context dataset [2].

2. Dataset with Thermal and Depth Images

The IPHD [1] dataset was compiled by researchers from the Chalearn Looking at People group. The entire set consists of over 100,000 pairs of images cut from two video streams, without maintaining information about their order. The footage was captured indoors and outdoors, at such places

Table 1

Detection evaluation metrics for single depth and thermal models based on the Faster R-CNN architecture computed based on the IPHD-test dataset. For reference, results of a method proposed by the authors of the IPHD database are shown as well. As no MAP was used in their publication, the relevant fields are marked NA (not available)

Method	Detection precision (\pm std. dev.)					
	Thermal images			Depth images		
	AP ₅₀	AP ₇₅	MAP	AP ₅₀	AP ₇₅	MAP
Single model (ours)	83.23% (± 0.28)	56.38% (± 0.32)	51.77% (± 0.29)	71.26% (± 0.59)	44.01% (± 0.45)	42.46% (± 0.49)
Baseline method from [1]	52.45%	15.95%	NA	34.32%	9.91%	NA

as streets, universities, libraries, and private houses. The video frames were spatially and temporally aligned. Therefore, they may be used independently or may be merged using one of the fusion methods. The extracted images were arbitrarily split into the development set (IPHD-dev), consisting of 84,818 images of both types, the validation set (IPHD-val), consisting of 12,974 images of both types, and the test set (IPHD-test), containing 15,115 images.

All the images are associated with ground-truth bounding boxes that show the position of people in the scenes. People visible in the images perform various actions: sitting on the sofa, lying on the floor, cooking, eating, talking on the phone. The scenes were manually labeled by the authors using RGB images. The RGB stream was also aligned with the other two, but it is not a part of the publicly available set. Manual labeling would be much more difficult or even impossible for depth or thermal images, since human body features are not easily distinguishable “with the naked eye”. One may notice (see Fig. 1) that the temporal alignment is imperfect, and some of the ground-truth detection boxes should be slightly shifted, especially the ones with individuals moving fast. This is caused by issues with the synchronization of sensors at a hardware level. Images in

the test set (IPHD-test) were manually adjusted to compensate for the misalignment. The effect of label weakness will also be discussed in this work. Because of the inferior human perception of alternative imaging sources, they are often considered privacy-preserving. However, the extent to which they do not contain any individual features needs to be investigated further.

2.1. Thermal Imaging

In thermal imaging, individual pixels represent far-infrared radiation measured by an IR detector. Radiation may be either emitted or reflected from the scene. The type of representation and its range vary across different sensors, but most thermal detectors have significantly lower resolution than DSC and video cameras used in other imaging systems. Images that are included in the IPHD dataset were collected with the FLIR Lepton 3 sensor. It is capable of detecting infrared waves with the length of 8 to 14 μm [3]. The original resolution of the sensor is 160 \times 120 pixels. The thermal images in IPHD were padded to ensure that their ratio is consistent with that of depth images. As a result, they are 1-channel 16-bit pictures with a resolu-

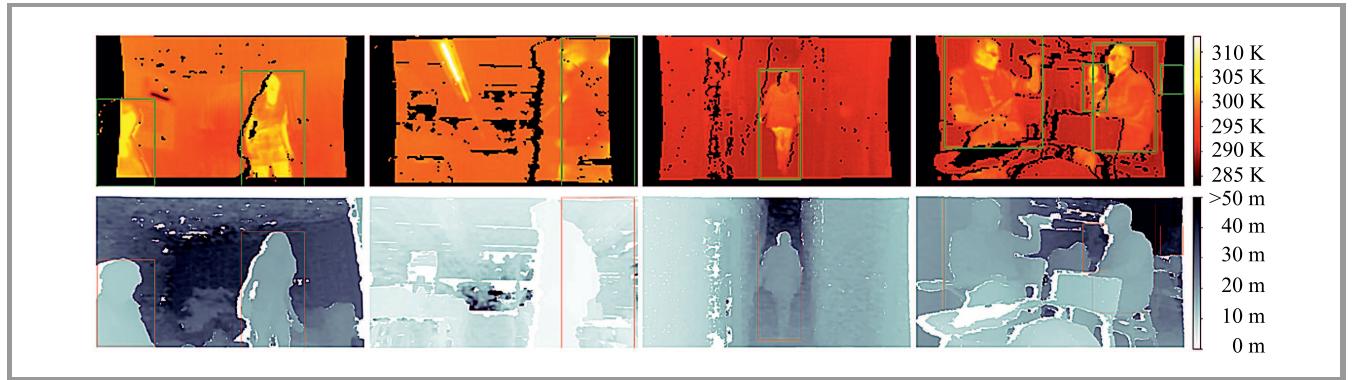


Fig. 1. Spatially and temporally aligned frames from the IPHD dataset: thermal images (upper row, in Kelvin) and depth images (bottom row, in meters). The images are shown in the altered color scales for better visualization (originally, one-channel 16-bit images). The bounding boxes indicate ground-truth locations of people (green boxes for thermal images, red boxes for depth images). Ground-truth labeling of non-RGB images can be challenging, because some of the features are difficult to perceive by the human eye. Here, in the IPHD dataset, labeling was performed using the corresponding RGB stream, which was not publicly available. Nevertheless, a temporal misalignment may be noticed in some of the bounding boxes (second column). The misalignment is caused by problems with synchronization between RGB and other sources. In the test part of the dataset (IPHD-test), the positions of boxes were adjusted manually. Therefore, this subset contains stronger labels than those in the training part (IPHD-dev). (see the digital edition for color images)

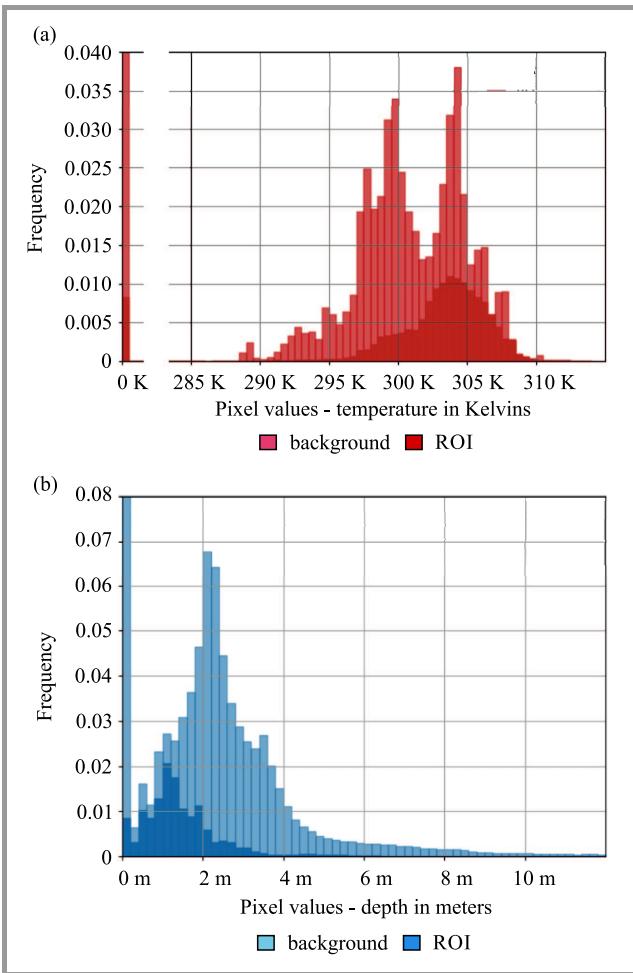


Fig. 2. Histograms of pixels from the IPHD dataset, computed separately for the areas occupied by human silhouettes (ROI) and for the background pixels. The temperature of pixels in thermal images (a) may vary from 0 to 450 K, with most values being within the 290–310 K range. Some of the pixels are set to 0 because of the spatial registration with depth images. Null pixels also denote depth data missing. The values of pixels in the depth images (b) range from 0 to 65 m, where 0 is typically caused by erroneous readings. Most people are positioned not further than 4 m from the camera.

tion of 213×120 pixels. Each pixel represents a specific level of temperature in the Kelvin scale. The padded portions of the images are filled with zeros, and zeros appear also at locations where the depth images render erroneous readings.

Temperature distribution of the images is presented in Fig. 2a. Histograms were computed independently for the areas occupied by human silhouettes and for the background. It may be observed that most of the pixels representing humans have temperatures in the range of 295 K to 310 K, with a mean value of approx. 303 K.

2.2. Depth Imaging

Depth cameras have become more popular and available due to the abundance of devices serving as game con-

trollers. Depth sensors may be easily used to map the pose of a human silhouette and other parts of the body, e.g. hands. Although depth imaging may be based on various hardware architectures, a typical consumer depth camera, like Microsoft Kinect v1 or Intel RealSense, consists of an active infrared projector and at least one infrared detector. The projector casts an invisible light pattern onto the scene. The distance of the object from the detector is estimated using the triangulation method that measures light pattern displacements [4].

The IPHD dataset contains depth images acquired with the use of the Intel RealSense D435 sensor. The size of the images equals 1280×720 pixels [5]. The images are 16-bit channels, just like their thermal counterparts. Pixel values represent the distance to the depth sensor, expressed in millimeters and have the maximum value of 65 meters. Null pixel values express erroneous readings. Bad pixels may appear at the borders of objects or human silhouettes and may also be caused by reflective materials or strong illumination.

Histograms of pixel values calculated for unprocessed depth images are shown in Fig. 2b. As far as the temperature is concerned, they were made separately for the background pixels and the pixels assigned to the human body. It may be observed that there are no labeled human subjects in the range greater than 4 meters: at these distances, the bounding boxes would be too small, and their contents would not be easily distinguishable.

2.3. Image Preprocessing

Before deploying the detection algorithms, the datasets were examined to choose the best preprocessing method. For the preliminary observations, a small subset of 50 images was drawn from IPHD-dev. It is hereinafter referred to as IPHD-pre. Using IPHD-pre, we manually labeled the masks that indicated precise people locations. The masks were used to select two distinct parts of the image: the foreground in which a person appears (ROI) and the background. The histograms of pixel values for those two subareas are shown in Figs. 2a and 2b, respectively. As the IPHD database authors suggested, the pixel values in thermal images should be standardized before further processing. In our work, the pixels in thermal images were clipped at the minimum value of $x_{min} = 285$ K and the maximum value of $x_{max} = 315$ K. Then, the images were normalized using the mean and standard deviation calculated on the full set of ROI pixels, excluding the null-value pixels. More precisely, each thermal image was preprocessed by extracting mean value $\tilde{x}_{th} = 296.4$ K and dividing it by $\sigma_{th} = 330$ K such as:

$$\begin{aligned} X[X > x_{max}] &= x_{max}, \\ X[X < x_{min}] &= x_{min}, \\ X &= \frac{X - \tilde{x}_{th}}{\sigma_{th}}. \end{aligned}$$

The depth images were also standardized before processing them with the use of the detection algorithms. Namely, they were clipped to match the range of $[0, \dots, 12 \text{ m}]$ and normalized with $\bar{x}_d = 2.830 \text{ m}$ and $\sigma_d = 3.198 \text{ m}$.

3. Human Detection Methods Applied to Depth and Thermal Images

There are several reasons why to include either depth or thermal imaging sources in the detection system. One is to augment the information about the objects for their further classification. The second is to leverage the precision of detection by adding potentially complementary imaging features. Finally, the third reason is to diminish the influence of bad lighting or image noise. The methods proposed in the literature are based on various combinations of image types: RGB and thermal [6], [7], RGB and depth [8], [9], [10] or, less commonly, thermal and depth [11].

This kind of a system is found to be functional in many real-world applications, from vision for autonomous driving [12], to industrial inspection [13], monitoring of car and vessel traffic [14], [15], drone surveillance [16], to pedestrian detection. In our work, we focus on the task of automatic person detection. Human detection solutions are widely discussed in the literature, but mostly in the context of RGB imaging. Most of the RGB-based methods apply also to alternative imaging sources. Older approaches are based on local descriptors, such as HOG or SIFT features [17], [18], which have to be hand-crafted. The recently proposed methods use predominantly convolutional neural networks inside their detection pipeline [19]. The most popular human detection architectures detection are Faster-RCNN [20], Mask-RCNN [21], SSD [22] or YOLO [23].

The authors of the IPHD database introduced the detection model based on YOLO in their paper [1]. A YOLO network is a single-stage detector that is available in various implementations, with its third version described in paper [23] still being one of the most commonly used variations. The IPHD baseline method for thermal-depth detection proposed a middle fusion network in which images are merged at the second-to-last convolutional layer level. The results generated by the algorithm serve as a point of reference for comparison with other methods, as they were computed under identical conditions and on the same dataset as in the case of our experiments.

The method proposed in our work is an extension of the Faster R-CNN method introduced in paper [20]. We have chosen this detection network to be the core of our system because it is easy to modify and perform better in terms of accuracy than YOLO (based on the results presented in [23]), as we do not intend to take into consideration other metrics, such as performance. Faster R-CNN involves a 2-stage detection procedure consisting of the region proposal network (RPN) and the second stage module respon-

sible for final object detection and classification. RPN produces a set of proposals that are then fine-grained. In the previous solution, named Fast-RCNN, these two stages were implemented by two different networks, but in Faster-RCNN, all functions are realized by one module. In our implementation, the Faster R-CNN network is built using the ResNet-50 [24] module (Fig. 3). The loss function has two components. One is the classification loss L_{cls} for assigning the probability of the object belonging to one of the classes. Here, as we perform solely person detection, the estimation distinguishes two classes only. The other is regression loss L_{reg} that compares the coordinates of box t_i with the ground-truth coordinates t_i^* where i is the box index, p_i is the measure of “objectness” and p_i^* is equal to 0 in the event of false detection.

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i * L_{reg}(t_i, t_i^*) \quad (1)$$

Typically, at the end of the detection procedure in R-CNN networks, some post-processing methods need to be applied to reduce the number of regions that overlap too closely. Non-maximum suppression (NMS) is one of the most commonly used algorithms, and its will be elaborated on in the section concerned with late fusion.

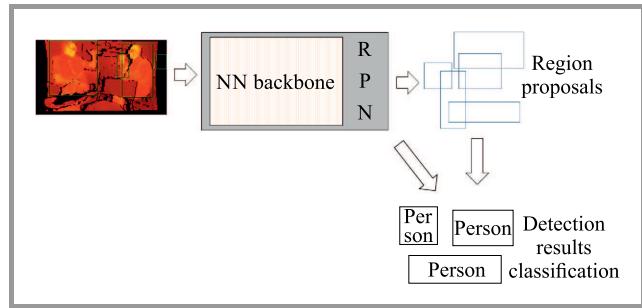


Fig. 3. Single detection model based on Faster-RCNN. Faster-RCNN typically consists of the RPN that generates a candidate list of detection boxes. RPN may be realized by one of the selected convolutional neural network architectures, such as ResNet-50 or AlexNet. At the end of data processing, ROI pooling is performed on the list of candidates to produce the final list of results with classification scores.

4. Results and Discussion

Average precision (AP) is typically used as the evaluation metric to test the detection methods. AP may be computed at different ranges of overlap between detection results and ground-truth bounding boxes. The level of precision with which two sets of coordinates are capable of describing the same object is measured by the intersection-over-union (IOU), defined by the area of overlap between two bounding boxes divided by the area of union.

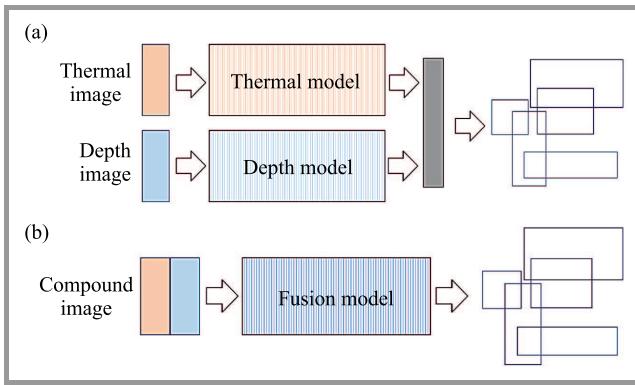


Fig. 4. We experimented with different model fusion patterns, namely late fusion (a), where the results of two models are merged into one set of results using dual non-maximum suppression, and early fusion (b) – where the images from thermal and depth streams are combined into one input of the model trained to recognize the compound input.

The authors of the IPHD dataset suggested AP₅₀ (average precision at IOU=50%) to be the primary evaluation metric of the challenge. This choice was motivated by the weakness of the ground-truth labels. Typically, in detection benchmarks, AP₇₅ is used, and it was calculated in our experiments as well (average precision at IOU = 75%). We also added MAP (mean average precision) as defined in the COCO challenge [2], being the mean value of AP at IOU=[50%: 5%: 95%].

Our experiments started by training two independent models based on Faster-RCNN networks, with ResNet-50 serving as the backbone (Fig. 4). The first one was trained on thermal images. Random crops and horizontal flips augmented the training set. Network weights were optimized with the use of the SGD method, with the learning rate

initialized at 0.005 and updated every three epochs. The thermal detector achieved AP₅₀ equal to 83.23% and AP₇₅ equal to 56.38% (see Table 1). All the results are computed on the test part of the IPHD dataset (IPHD-test) and are averaged for three repetitions of the training and evaluation procedures.

The second model – with its architecture identical to the first one – was trained on depth images. For this setup, we obtained AP₅₀ = 71.26% and AP₇₅ = 44.01%, respectively. Detection evaluation metrics for the depth model were much lower than those for the thermal network. Therefore, we may consider the depth data to be more challenging than their thermal counterparts for running the detection algorithms on.

Some qualitative results of experiments with single-model detection are shown in Fig. 5. We can see that for the depth model, there is often a more significant number of false detections (third column in Fig. 5) and missed detection, especially for smaller objects (second column in Fig. 5). Detection errors may be easily explained, since some objects are difficult to distinguish from people in the depth scenes for the human observer. On the other hand, we noticed false detections pointing to a dog visible in the thermal images. It would be probably reduced by adding labels of other warm objects (like animals and electrical appliances) to teach the network to distinguish them from humans.

4.1. Fusion of Models

Fusion of deep neural models is a technique that can typically boost the accuracy of results, as shown in paper [25] in relation to the classification of videos. The authors distinguished three approaches to classify the content of multi-frame data: early fusion, late fusion, and slow fu-

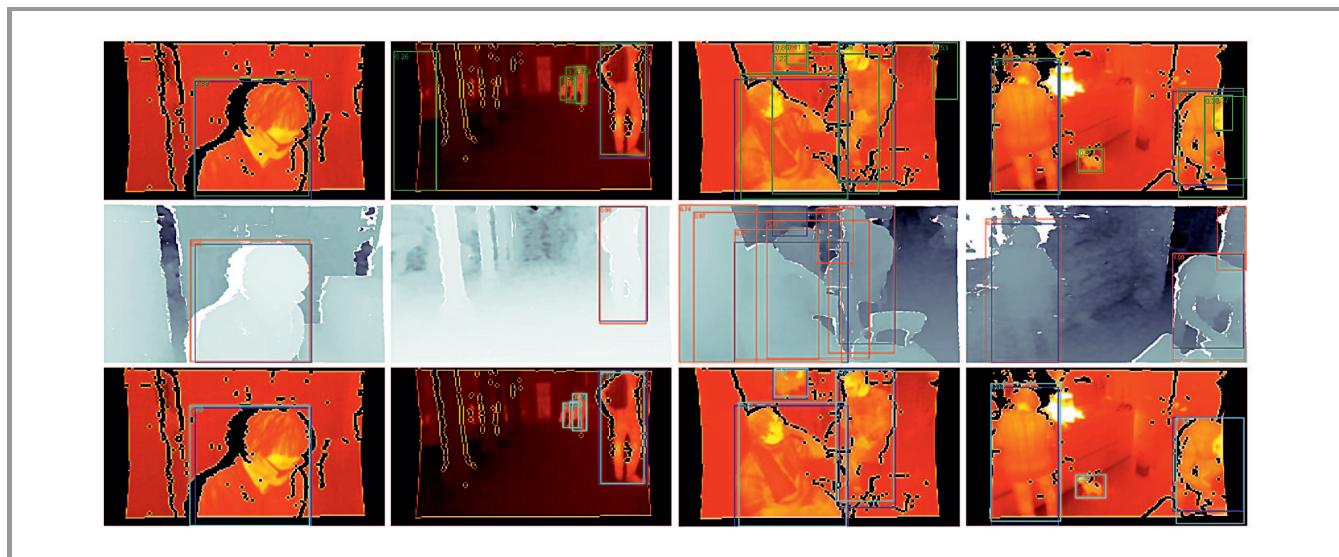


Fig. 5. Qualitative human detection results using IPHD-test data. Blue rectangles indicate ground truth bounding boxes. The consecutive rows show the detections for a single thermal model (first row, green boxes), the detections for a single depth model (second row, red boxes), and predictions from the model with early fusion (third row, cyan boxes).

sion. The division is defined by the point of the network at which the flow of information is merged. Early fusion means concatenating the data before further processing it inside a model. Consequently, in late fusion, the data is merged at the end of the model. In slow fusion (or middle fusion), merging is performed at some of the intermediate layers of the model. The slow fusion technique was applied to the detection network in the IPHD method discussed here for reference (results showed in Tab. 1). Also, the two other approaches (late and early fusion) can be transferred directly to the domain of multimodal images. Therefore, we decided to employ them in our experiments. Late fusion was realized by the application of different versions of the non-maximum suppression algorithm (NMS). Models trained in the previous experiments were reused as components of the fusion system. In the case of early fusion, it was necessary to change the structure of the model and perform the entire training procedure. Results for the fusion strategies were compared against results for the single image approaches discussed in the preceding section. The code of the library was written with PyTorch and it is available online: github.com/weronikagutfeter/Red-Hot-Deep-Blue.

4.2. Dual-NMS

Our late fusion method was based on non-maximum suppression (NMS) [26]. NMS is a post-processing algorithm that is typically used to minimize the number of redundant and overlapping detection results. Simple non-maximum suppression, also called greedy NMS, begins with sorting the detection boxes by their scores in the descending order. Then, results from the sorted list are compared, one by one, with the remaining results. If the IOU of the compared pair of boxes is larger than a selected threshold value, the box with the lower confidence score is removed.

This paper employs a modified version of the NMS algorithm to merge the detection boxes originating from the two distinct models: thermal and depth. The modified version

of the algorithm is called dual-NMS. The idea behind this approach is to collect pairs from the two lists of detection boxes, which are also sorted by their confidence scores, like in the simple NMS method. Boxes with the highest scores (from the depth or the thermal lists) are taken, one by one, and compared with all boxes from the other list. The selected detection box is paired with the result with a sufficient IOU and the highest score among the candidates from the other list. The pair is then merged into a single result, and the final detection box coordinates are updated by applying weighted averaging of the coordinates of the components.

Since some unassigned detection boxes may be left after the pairing, several approaches to managing unpaired boxes were evaluated. The simplest solution is to remove the unassigned results, as they are either not present in any stream, have low confidence scores, or are duplicates. However, one of the streams is likely to be a more robust source of detection results. Thus, we may leave the unpaired samples from this source.

To sum up, we checked four versions of the algorithm: with all of the unpaired results removed, with all of the unpaired results kept, with only the unpaired thermal results kept, and, finally, with the unpaired depth results kept. For the reference, we compared dual-NMS with the simple NMS algorithm applied to the concatenated list of detection boxes. Precision rates of the fusion are presented in Table 2. For comparison, the results for single model detectors from Table 1 are shown at the bottom. The best results for late fusion approaches were obtained for the dual-NMS with thermal results kept, for which $AP_{50} = 83.31\%$ and $AP_{75} = 57.84\%$. However, this model was only slightly better than the single thermal model. It can be concluded that for the NMS-based system, the impact of the depth detection module is relatively low.

Table 2

Average precision of detection computed on the IPHD-test dataset for Faster-RCNN after applying a priori (early fusion) and a posteriori (late fusion) merging techniques. Two best results – one for early fusion and one for late fusion are shown in bold print

Fusion strategy	Detection precision (std. dev.)			
	Thermal + depth images			
	AP_{50}	AP_{75}	MAP	
Late fusion with Dual-NMS	Leave all unpaired	77.28% (± 0.72)	55.14% (± 0.48)	49.91% (± 0.57)
	Leave unpaired thermal	83.31% (± 0.39)	57.84% (± 0.14)	53.18% (± 0.32)
	Leave unpaired depth	74.90% (± 0.56)	50.43% (± 0.45)	47.07% (± 0.54)
	Remove all unpaired	69.63% (± 0.44)	52.98% (± 0.14)	46.76% (± 0.33)
Late fusion with Simple-NMS		73.26% (± 0.38)	51.39% (± 0.68)	46.62% (± 0.47)
Early fusion		88.86% (± 0.19)	63.82% (± 0.21)	57.42% (± 0.38)
Single model	Thermal	83.23% (± 0.28)	56.38% (± 0.32)	51.77% (± 0.29)
	Depth	71.26% (± 0.59)	44.01% (± 0.45)	42.46% (± 0.49)

4.3. Early Fusion

The early fusion approach requires modification of the model structure in order to accept concatenated data sources serving as input for the network. The proposed method encodes a pair of images consisting of one thermal image and one depth image in a single pass of the network. The images are rescaled to 1280×720 pixels to ensure their consistent size, and are then stacked to create a 2-channel multimodal image. As the model structure is less typical than the structures commonly used in Faster R-CNN networks, it limits the options of transfer learning and requires full retraining. The training dataset needs to be adequately prepared. In the experiments, both channels were standardized in the same way as in single-image detection. For a fair comparison, the fusion model was also realized using Faster R-CNN architecture with ResNet-50 backbone, as it was the case in the previous tests.

The early fusion approach surpassed all other methods tested in this paper. The precision rates obtained equaled $AP_{50} = 88.86\%$ and $AP_{75} = 63.82\%$, respectively. The result was better than for the single thermal model by 6.7% (AP_{50}) and 13.2% (AP_{75}), and surpassed the dual-NMS method by 6.6% (AP_{50}) and 10.3% (AP_{75}). Qualitative results for the early fusion model are shown in the third row of Fig. 5. Some of the missing and false detections were eliminated as a result of applying fusion to the detector.

5. Conclusions

The experiments described in this paper prove that it is possible to detect humans, both in thermal and depth images, while achieving acceptable precision rates. The rates are acceptable but still far from the precision level that may be obtained for RGB images. When comparing the two types of streams, it may be noticed that the images with information about the temperature are a better source of visual information for identifying people. In depth images, people are harder to distinguish from other objects, both for our algorithms and for humans. On the other hand, the thermal network has more false positives indicating other warm objects, such as animals or electric equipment. This is not the case in the depth model. We show that merging two types of imaging sources is capable of improving the outcomes generated by the detection network. We tested two important fusion strategies: early fusion (combining images at the network's input) and late fusion with a modified non-maximum suppression algorithm, namely dual-NMS. Both variants showed improvements in comparison to single-model detection. The best solution was obtained when the model was retrained on compound images (early fusion), and the result was better than the one achieved with the dual-NMS approach. However, we must stress that preparing early fusion models requires more effort and computational resources. We used an aligned dataset to make the training procedure possible. This type of data is not always available. Late fu-

sion approaches allow merging the detection results from two independent single-source models trained on unaligned data. Further work on the methods under consideration requires the dataset to be extended. Access to an RGB source, for reference purposes, could be valuable for the development of the algorithm. The order of frames is another piece of information that is missing but may be obtained. After proper labeling, the detection method can be developed further and converted into a human tracking solution.

References

- [1] A. Clapés, J. S. Jacques Junior, C. Morral, and S. Escalera, “Chalearn lap 2020 challenge on identity-preserved human detection: Dataset and results”, in *15th IEEE Int. Conf. on Automatic Face and Gesture Recogn. (FG 2020)*, Buenos Aires, Argentina, pp. 859–866, 2020 (DOI: 10.1109/FG47880.2020.00135).
- [2] T.-Y. Lin *et al.*, “Microsoft COCO: common objects in context”, CoRR, abs/1405.0312, 2014 [Online]. Available: <https://arxiv.org/pdf/1405.0312>
- [3] Teledyne Flir LLC, “LWIR micro thermal camera module Lepton 3”, 2018 [Online]. Available: <https://www.flir.com/products/lepton/?model=500-0276-01> (accessed on: 01.01.2021).
- [4] J. Smisek, M. Jancosek, and T. Pajdla, *3D with Kinect, Consumer Depth Cameras for Computer Vision. Adv. in Computer Vision and Pattern Recogn.*, A. Fossati, J. Gall, H. Grabner, X. Ren, K. Konolige, Eds., pp. 3–25. Springer London, London: 2013 (DOI: 10.1007/978-1-4471-4640-7_1).
- [5] Intel Corporation, Intel RealSense Depth Module D400 Series CustomCalibration, 2019 [Online]. Available: https://www.intel.com/content/dam/support/us/en/documents/emerging-technologies/intel-realsense-technology/RealSense_D400%20_Custom_Calib_Paper.pdf (accessed on: 01.01.2021).
- [6] S. Kumar, T. K. Marks, and M. Jones, “Improving person tracking using an inexpensive thermal infrared sensor”, in *IEEE Conf. on Computer Vision and Pattern Recogn. Workshops*, Columbus, OH, USA, pp. 217–224, 2014 (DOI: 10.1109/CVPRW.2014.41).
- [7] A. S. Charan, M. Jitesh, M. Chowdhury, and H. Venkataraman, “Abifn: Attention-based bi-modal fusion network for object detection at night time”, *Electronics Letters*, vol. 56, no. 24, pp. 1309–1311, 2020 (DOI: 10.1049/el.2020.1952).
- [8] H. Haggag, M. Hossny, S. Nahavandi, and O. Haggag, “An adaptable system for RGB-D based human body detection and pose estimation: Incorporating attached props”, in *IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, Budapest, Hungary, pp. 1544–1549, 2016 (DOI: 10.1109/SMC.2016.7844458).
- [9] O. H. Jafari, D. Mitzel, and B. Leibe, “Real-time RGB-D based people detection and tracking for mobile robots and headworn cameras”, in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, Hong Kong, China, 2014, pp. 5636–5643 (DOI: 10.1109/ICRA.2014.6907688).
- [10] M. Rasoulianesh, S. Yadav, S. Herath, Y. Vaghei, and S. Payandeh, “Deep attention models for human tracking using RGBD”, *Sensors*, vol. 19, no. 4, 2019 (DOI: 10.3390/s19040750).
- [11] H. S. Hadi, M. Rosbi, U. U. Sheikh, and S. H. M. Amin, “Fusion of thermal and depth images for occlusion handling for human detection from mobile robot”, in *10th Asian Control Conf. (ASCC)*, Kota Kinabalu, Malaysia, pp. 1–5, 2015 (DOI: 10.1109/ASCC.2015.7244722).
- [12] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, “Sensor and sensor fusion technology in autonomous vehicles: A review”, *Sensors*, vol. 21, no. 6, 2021 (DOI: 10.3390/s21062140).

- [13] V. F. Vidal *et al.*, “Sensors fusion and multidimensional point cloud analysis for electrical power system inspection”, *Sensors*, vol. 20, no. 14, pp. 40–42, 2020 (DOI: 10.3390/s20144042).
- [14] T. Alldieck, C. H. Bahnsen, and T. B. Moeslund, “Context-aware fusion of RGB and thermal imagery for traffic monitoring”, *Sensors*, vol. 16, no. 11, 2016 (DOI: 10.3390/s16111947).
- [15] F. Farahnakian and J. Heikkonen, “Deep learning based multi-modal fusion architectures for maritime vessel detection”, *Remote Sensing*, vol. 12, no. 16, 2020 (DOI: 10.3390/rs12162509).
- [16] A. Morfin-Santana *et al.*, “Real-time people detection from thermal images by using an unmanned aerial system”, in *16th Int. Conf. on Electric. Engineer., Comput. Sci. and Automatic Control (CCE)*, Mexico City, Mexico, pp. 1–6, 2019 (DOI: 10.1109/ICEEE.2019.8884561).
- [17] S. Chang, F. Yang, W. Wu, Y. Cho, and S. Chen, “Nighttime pedestrian detection using thermal imaging based on hog feature”, in *Proc. 2011 Int. Conf. on System Sci. and Engineer.*, Macau, China, pp. 694–698, 2011 (DOI: 10.1109/ICSSE.2011.5961992).
- [18] L. Spinello and K. O. Arras, “People detection in RGB-D data”, in *IEEE/RSJ Int. Conf. on Intell. Robots and Systems*, San Francisco, CA, USA, pp. 3838–3843, 2011 (DOI: 10.1109/IROS.2011.6095074).
- [19] Ch. Herrmann, M. Ruf, and J. Beyerer, “CNN-based thermal infrared person detection by domain adaptation”, in *Proc. Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, M. C. Dudzik and J. C. Ricklin, Eds., Orlando, FL, USA, vol. 10643, 2018, pp. 38–43 (DOI: 10.1117/12.2304400).
- [20] Sh. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks”, in *Proc. of the 28th Int. Conf. on Neural Information Process. Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 1, pp. 91–99, 2015 [Online]. Available: <http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf>
- [21] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, “Mask R-CNN”, CoRR, abs/1703.06870, 2017 [Online]. Available: <https://arxiv.org/pdf/1703.06870>
- [22] W. Liu *et al.*, “SSD: single shot multibox detector”, CoRR, abs/1512.02325, 2015 [Online]. Available: <https://arxiv.org/pdf/1512.02325>
- [23] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement”, CoRR, abs/1804.02767, 2018 [Online]. Available: <https://arxiv.org/pdf/1804.02767>
- [24] K. He, X. Zhang, Sh. Ren, and J. Sun, “Deep residual learning for image recognition”, CoRR, abs/1512.03385, 2015 [Online]. Available: <https://arxiv.org/pdf/1512.03385>
- [25] A. Karpathy *et al.*, “Large-scale video classification with convolutional neural networks”, *27th IEEE Conf. on Computer Vision and Pattern Recogn. (CVPR)*, Columbus, OH, USA, 2014 (DOI: 10.1109/CVPR.2014.223).
- [26] A. Neubeck and L. Van Gool, “Efficient non-maximum suppression”, in *18th Int. Conf. on Pattern Recogn. (ICPR'06)*, Hong Kong, China, vol. 3, pp. 850–855, 2006 (DOI: 10.1109/ICPR.2006.479).



Weronika Gutfeter is a research assistant at the Biometric and Machine Intelligence Laboratory of the Research and Academic Computer Network (NASK) in Warsaw, Poland. She obtained her Master's degree in Computer Engineering from Warsaw University of Technology and is currently working on her Ph.D. thesis

that covers the topics of 3D, semi-3D and multi-view face identification methods. Her research interests include also other computer vision and biometric identification methods, such as off-line signature verification and iris detection. At NASK, she works on developing machine learning systems for the public sector.

<https://orcid.org/0000-0001-6359-8220>

E-mail: weronika.gutfeter@nask.pl

Biometric and Machine Intelligence Laboratory
Research and Academic Computer Network (NASK)
ul. Kolska 12
Warsaw, Poland



Andrzej Pacut is a full Professor, chair of the Biometrics and Machine Learning Group at the Faculty of Electronics and Information Technology in Warsaw University of Technology. He is the founder of the Technical Committee on Biometrics (mirror of ISO/IEC JTC1 SC37) within the Polish Standardization Committee (PKN),

chair of the Biometrics and Machine Intelligence Laboratory at the Research and Academic Computer Network (NASK), and a member of the NASK Research Council.

<https://orcid.org/0000-0003-3489-8990>

E-mail: andrzej.pacut@nask.pl

Biometric and Machine Intelligence Laboratory
Research and Academic Computer Network (NASK)
ul. Kolska 12
Warsaw, Poland