Depth and Thermal Images in Face Detection - A Detailed Comparison Between Image Modalities

Wiktor Mucha
wiktor.mucha@tuwien.ac.at
TU Wien, Computer Vision Lab
Vienna, Austria

Martin Kampel
martin.kampel@tuwien.ac.at
TU Wien, Computer Vision Lab
Vienna, Austria

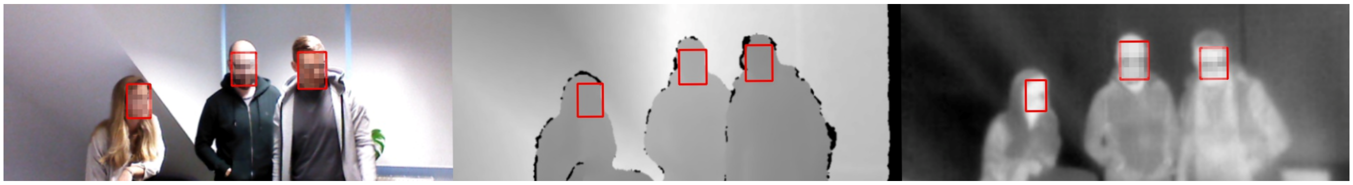

Fig. 1: Example of correct face detection in each image: RGB, depth, thermal.

ABSTRACT

Face detection is a well-known issue in image processing, and numerous studies are present in this field. A prominent part of the work is devoted to RGB images, leaving depth and thermal data with less interest. However, in some conditions like low-light areas where face detection is needed, non-RGB sensors might perform better. Also, mounting an additional RGB camera could be challenging or not possible, considering privacy concerns. In this work, current deep learning methodologies are employed to train depth and thermal detection models. The training is done using combined publicly available data that is processed by us for this purpose in order to create necessary annotations for a learning process. The resulting models are validated on a new trimodal dataset collected for this experiments purpose. It contains images captured with RGB, depth, and thermal sensors. Various scenes with single and multiple faces appearances can be found. The results show that non-RGB solutions can be applied in practice with highly robust accuracy and their efficiency is close to RGB detectors. However, their performance depends on the environment and that circumstances are described later in this article.

CCS CONCEPTS

• **Computing methodologies → Object detection**.

KEYWORDS

image modalities, depth face detection, thermal face detection, datasets, modalities comparison, deep learning

ICMVA 2022, February 18–20, 2022, Singapore, Singapore
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9567-0/22/02…$15.00
https://doi.org/10.1145/3523111.3523114

ACM Reference Format:
Wiktor Mucha and Martin Kampel. 2022. Depth and Thermal Images in Face Detection - A Detailed Comparison Between Image Modalities. In *2022 the 5th International Conference on Machine Vision and Applications (ICMVA) (ICMVA 2022), February 18–20, 2022, Singapore, Singapore.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3523111.3523114

1 INTRODUCTION

Face detection is one of the fundamental topics in the area of computer vision which has been studied for a long time now. These techniques have become usual in our life thanks to smartphones that we use daily. They are implemented in various applications available on our mobiles. For example, face detection is a part of the face recognition pipeline[18] where it can be exploited for phone unlocking. However, most of the algorithms in this field use RGB data or a combination of RGB and depth images for better efficiency. Using a single depth sensor or thermal imaging has not been studied widely in this field. There is no existing comparison between these three techniques in face detection or any benchmark available publicly. In practical applications of vision systems, face detection or face tracking may be needed, but the applied system has only one type of camera available, which is not RGB type. For example, systems that are tracking human pose and body. They might be implemented in car interiors as driver assistance like in work of Silva et al.[23]. Another case is Active and Assisted Living (AAL) technologies for the elderly who require specific medical care. Depth or thermal data brings fewer privacy concerns like in Banerjee et al.[4] study. Is it necessary to extend such a system with an additional RGB camera, or is available depth or thermal sensor enough to perform correct face detection? How do the depth and thermal detections perform in comparison to the RGB data approach? In this work, to answer these questions, we implement state-of-art Convolutional Neural Networks (CNN) for face detection tasks on three different imaging sensor types: depth, thermal and RGB. We prepare the training data for depth and thermal models from five publicly available datasets, including annotation creation, and perform algorithm learning processes. In the case of the RGB detector, the model used is already trained. Finally, our models are evaluated on a new dataset gathered by us that contains exactly the same scenes, but captured with three different camera types mounted together. There are 2808 RGB, depth, and thermal samples including both single and multiple faces occurrences. The images vary in backgrounds, scenes, persons, and lighting conditions. Depending on the scenery complexity they are divided between *easy*, *medium* and *hard* subsets.

The main contribution of this paper includes:

- Creation of annotations for faces in already existing public datasets for training purposes. The annotations are planned to be publicly released.
- Introduction of a new trimodal dataset named TriModal Face Detection(TMFD) Dataset for face detection task planned to be publicly released. The novelty of the dataset is not only in various modalities but in perfectly aligned images of the same scene captured by depth, thermal and RGB sensors. The dataset might be a benchmark in the field of face detection from depth or thermal data and allows direct comparison between these modalities.
- Application of CNN for face detection to depth and thermal images.
- Performance study and comparison between these three modalities in an example of face detection task.

This paper is organized as follows. Section 2 reviews related work in the field of face detection from the depth and thermal images. Section 3 describes the chosen methodology with justification. The applied deep learning architecture is presented. The last subsection is devoted to the data preparation process of training and validation samples for depth and thermal models. Section 4 is focused on the experiment. It includes clarification of the test dataset preparation with a detailed description of the dataset. The metrics used are provided with an explanation. Finally, results are presented and discussed. Section 5 summarizes this paper by highlighting the key observations from the carried study.

2 RELATED WORK

In the field of face detection, all three data types of color, depth, and thermal are applicable. Typically, depth data appears in the context of improving the performance of algorithms based on RGB images as in works[17][19][20]. Frameworks using only depth or only thermal samples are less common than these based on RGB inputs.

2.1 Face detection from depth data

One of the first works in face detection from depth data only is presented by R. Hg et al.[11]. The authors contribute to the topic with a new methodology and an RGB-D dataset for face detection and recognition. The algorithm finds the closest object to the camera to reduce the search space. Then the curvature analysis technique searches for face candidates. These measures are used for classification to determine the type of surface. Detected surfaces are merged into triangles of structure eye-eye-nose. Face image is transformed to become faceness, it is centered, masked, and transformed. Later on, it is validated using the Principal Component Analysis method. The face space is built. The testing image is projected into this face space and then back to the original image space. If the reconstruction error fits the result threshold, it is considered as a face. More recently, feature-based studies have been presented. One of these examples is the work of Li et al.[14]. Depth data is mapped into a 2d image combined with a smooth image processing method to get depth images. An improved HOG-LBP algorithm using Histograms of oriented gradient (HOG) and local binary patterns (LBP) is designed to profile the features of a face depth. Classification is done by Support Vector Machine (SVM). The feature-based methodology is presented as well in the work of [25]. Authors employ Haar-like features as a descriptor of images. Then a Bilateral filter is applied, the foreground is distinguished from the background to obtain a face mask. They compare their implementation with traditional 2d based frameworks in different lighting conditions. They outperform RGB detection in low-light environments where there is not enough light to reconstruct the scene correctly with an RGB sensor, and information for algorithm processing is missing.

A similar task to face detection is the head detection issue. The main difference is in a more considerable range of camera angles accepted as true candidates than faces, making this a more challenging problem. In the case of faces, characteristics are eyes, ears, or nose. In head detection, none of these characteristic points may be visible for the sensor. For the stated reasons, the topic of head detection has been researched too as an extended case of face detection.

A head detection solution is presented by Chen et al.[6] using their own, novel head descriptor. It classifies pixels as head center or non-head center. A false-positive filter is employed to select the most probable head centers with clustering to determine final locations. The algorithm outperforms HOG-based detectors. Depth-only head detection is also presented by Hacinecipoglu et al.[10]. The framework removes the ground plane from images. Further, they apply filtering and clustering, which leads to extracting human bodies from scenes. The final stage is head area extraction and classification performed by the SVM algorithm. The whole pipeline is designed to work on a mobile robot. Deep learning (DL) techniques are equipped for head detection in studies of Diego Ballotta et al.[3][2]. They present a network that outperforms other solutions at that time. In the second study, they introduce improvements to the architecture and overcome the sliding-window approach with Fully Convolutional Network. They train and validate the model on two datasets. Their algorithm exceeds state-of-art methods and is able to run in real-time. However, it is not evaluated on images that contain multiple heads appearances in one scene.

2.2 Face detection from thermal data

One of the first face detection applications from thermal data is presented by Cheong et al.[7]. It is based on Otsu's thresholding method for converting thermal images to binary form. The horizontal projection is calculated for the image to identify the global minimum. It helps to identify the height and width of the head region. They also present their custom database for final verification. The research of Kopaczka et al.[13] is focused on adapting machine learning (ML) algorithms used for RGB data to the thermal datasets. ML approaches are described as proven to outperform traditional detections based on thermal information. Comparison is made between several image descriptors. The best performance is achieved with the Deformable parts model (DPM), HOG, and LBP extractors. The conclusion is that a proper dataset is needed to perform training correctly. In the work of Silva et al.[23] the authors explore the usage of a thermal camera to detect faces of persons in a car interior. They use DL with a YOLOv3 detector for RGB images. They apply transfer learning with a model pre-trained on RGB images and consider full retraining in feature work. The biggest constraint is in the sensors range, which is only available to perform up to 60 cm. Another study with a similar approach is presented by Vukovicet al. l[24]. This time, the model used is named Region-based Convolutional Neural Network (R-CNN), and transfer learning is applied. It performs face detection from thermal images only. Authors evaluate it on the dataset collected by them. What is unique, it consists of multiple faces occurring in single scenes. However, computational time does not allow to perform this application in real-time. The authors suggest using the Faster R-CNN or YOLO algorithm.

3 METHODOLOGY

Considering the works cited in the previous section, we apply a solution based on deep learning. The work of Ballotta et al.[2] in the head detection field uses a CNN network where VGG-19 network is used as a feature extractor. This shows that feature extractors designed for RGB data can adapt to depth samples if trained from scratch. Similar conclusions are coming from the analysis of thermal studies. In experiments referenced in [23] and [24] it is proven that CNN are able to perform face detection on thermal images. However, in both works, feature extraction parts of the model are pre-trained on RGB data. Due to this fact, we decide to retrain the existing state-of-art model for face detection on depth and thermal data. Nevertheless, due to the lack of existing datasets, we create our training image sequence from existing sets available in public.

3.1 Face detection model

In this work the *scrfd_1g* model is implemented. It belongs to a family of SCRFD DL architectures for face detection presented by Guo et al.[9]. They achieve top performance on WIDER face[26] benchmark with *scrfd_34g* having one of the best AP results. However, analysing the results and performance of the whole family of models, the *scrfd_1g* is chosen as a compromise between computational speed and achieved scores.

The *scrfd_1g* is built from three common parts in the field of object detection algorithms, a backbone, a head, and a neck. The first part is made out of MobileNetV1[12], a network designed for embedded and mobile devices. It is responsible for learning and extracting features from images. For the neck part Path Aggregation Feature Pyramid Network (PAFPN)[16] is used. This network runs instance segmentation and forms features from a previous network part by pooling them from all levels. It shorten the distance among lower and topmost feature levels for robust information transfer. Extracted features are passed to the head stage, and outputs are calculated. The head consists of stacked 3 × 3 convolutional layers. In this part of network classification loss is retrieved using Generalised Focal Loss (GFL)[15] and regression branches loss for bounding boxes with DIoU loss[27].

For both depth and thermal models, identical parameters during their learning process are used. The parameters are learning rate equal 0.01, momentum equal 0.9, and weight decay is 0.0005. Each sample is appropriately normalized to have a mean value equal to 127.5 and a standard deviation equal to 128. Further, images are randomly cropped, resized, and flipped. The training is run for 640 epochs with a batch size equal to 8.

3.2 Dataset preparation

There are fewer datasets available for face detection from depth images than in the RGB field. In the existing ones, images are usually taken in laboratory conditions, where a single person sits in front of a wall. This differs from popular RGB face detection benchmarks like WIDER face[26] where in some cases, hundreds of people are observed on one captured scene in real-world conditions. The goal is to evaluate a presented algorithm on data closer to real-world conditions with multiple appearances of faces, as opposed to available studies. Our depth dataset is created from four different sets publicly available. Images from BIWI[8] which contain segmented depth masks of single persons in the scene are selected. Another sequence is taken from Pandora[5] where the scene is similar to BIWI[8], but without background segmentation. To differentiate data, images containing real-world environment with multiple face occurrences from AAU VAP Trimodal People[21] and Hallway RGBDT[1] datasets are added too. All of the used datasets are captured as image sequences. Because of this, repeated or similar samples are deleted. In the end, our training set consist of 1294 samples and validation of 354 samples.

In the case of thermal images, circumstances are similar. Most of the available images are already processed for face recognition or are not challenging enough for the object detection task. Thermal data for face detection has the same problem with a availability like in the field of depth images. There are only a few existing datasets for this task, and it makes training preparation more difficult. In this study, images from the already mentioned AAU VAP Trimodal People[21] dataset that includes thermal images as well and The Tufts Face Database[22] are merged. As a result, the proportion between validation and training samples for learning the thermal model equals 660 to 1814 samples.

All of the mentioned and used data does not have available annotations for the face detection task. Due to this fact, labels are created manually by applying the *scrfd_34g* model, which is the most effective from the whole SCRFD family. Every generated ground truth is checked, and the annotations are added or corrected. Created annotations are planned to be released publicly for further use.

4 EXPERIMENT AND RESULTS

4.1 Experiment

We evaluate the performance of three models on our test data. Images are captured by a depth sensor Orbbec Astra and Flir Lepton 3.5 thermal camera. The first device is responsible for gathering samples in RGB color space and 16bit depth data. The second one collects 16bit thermal data. Both cameras have been mounted together on a tripod to produce exactly the same frame. Images contain scenes with single and multiple visible faces.

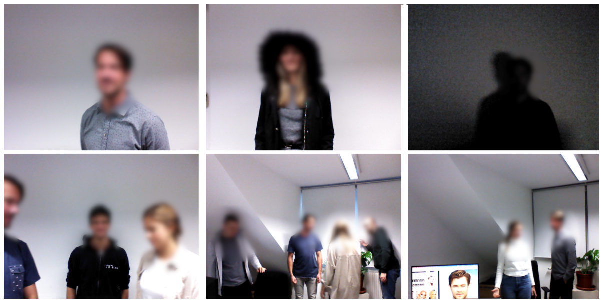

Fig. 2: Examples from our *easy*, *medium*, *hard* subsets

4.2 TMFD Dataset introduction

Due to the lack of available benchmarks in the field of depth and thermal face detection, the creation of our own dataset is needed. However, there are no existing trimodal datasets for this task, including depth, thermal and RGB images. Because of these circumstances, a new dataset is collected in this study to perform a valid comparison, which is planned to be made publicly available. Each RGB image in this dataset has corresponding frames of depth and thermal data. An example of three aligned frames from different sensors is presented in Fig. 1. This dataset contains images with different variations. The samples vary in the number of persons present in the scene, the individual persons, the type of background, the measurement distance, the wearable accessories used as obstacles (hoodies, headphones, hats, glasses, face masks), and the type of illumination in the scene. The images are categorized into three separate groups based on their complexity and difficulty level in the context of face detection. All samples presented in this paper are blurred only for the purpose of this document for privacy concerns.

The first subset contains the easiest detection conditions. This category covers images in which a single person is placed against a simple background. The sensor is positioned at a close range from the target. An example image is shown in Fig. 2(1). Further diversity is in the variation of lighting conditions in the captured scene, presented in Fig. 2(3). This subset contains 781 images of each modality.

The second batch incorporates medium detection conditions. In this group, the sensor placement remains the same as in the easier subgroup, but additional variety is introduced. Individuals are captured with wearable accessories as obstacles to make the correct detection harder (Fig. 2(2)). These accessories are hoodies, headphones, hats, and glasses. The second group includes the appearance of multiple persons in the scene to test the multiple detection performance presented in Fig. 2(4). In this section, there are 965 images of each modality.

The last group takes the most difficult images that remain. This time, the sensor is placed further away from the target. The scene features various obstacles, including computer screens, plants, desks, and chairs. The persons in the scene interact with them (Fig. 2(5)). Their occurrence is both singular and plural in the scene. Some of the samples include false-positive faces projected on a computer screen like the one presented in Fig. 2(6). A distinct subset in this collection are images from a previous, easier scene, but with a face mask usage to increase the difficulty of the prediction task. This particular group contains 1062 images of each modality.

Table 1: Comparison of a performance in the presented dataset for each modality.

Average precision for three image types AP[%]	Easy (781 samples)	Medium (965 samples)	Hard (1062 samples)
RGB	99.99	99.99	92.95
Depth	97.91	96.55	77.04
Thermal	98.98	94.07	83.98

4.3 Metrics

To evaluate the models, we employ standard metrics in the field of object detection. These are precision, recall, and average precision (AP).

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

Where:

- tp: is the number of correct face detections
- fp: is the number of correct face rejections
- fn: is the number of incorrect face rejections

AP is measured over various threshold values of prediction confidence score for each bounding box. The Intersection over Union (IoU) is fixed. There are 1000 threshold values between 0 and 1 in our calculations. The final value of AP is calculated as an area under the precision-recall curve of each plot presented.

4.4 Results

All of the three models are able to perform face detection on our dataset. The difference between performance is visible between three subsets of the test data. The results are shown in a Table 1.

In the *easy* subset, all three models perform on a similar level where RGB model leads scores with AP of 99.99%, followed by thermal solution with AP of 98.98%. The depth model presents the

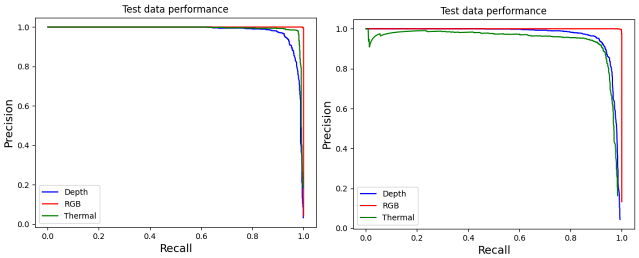

Fig. 3: Precision-recall curves for modalities in the *easy* (left) and *medium* (right) data subsets.

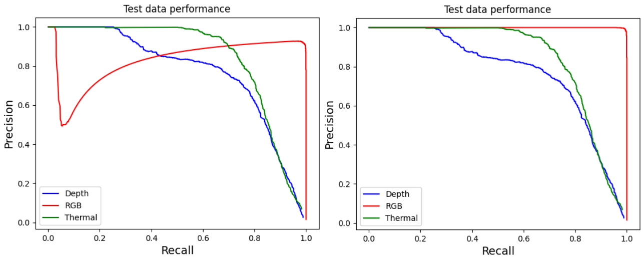

Fig. 4: Precision-recall curves for modalities in the *hard* data subset with false positive faces (left) and without (right).

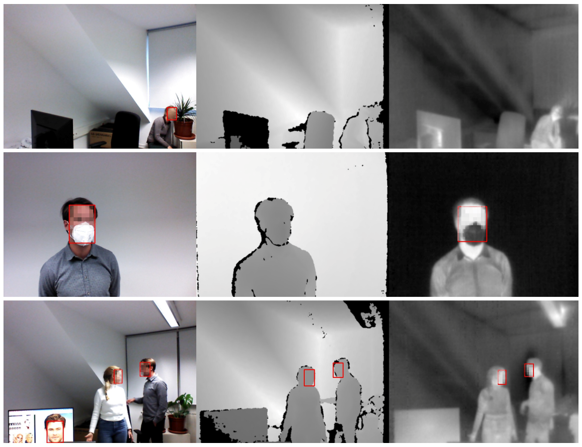

Fig. 5: From the top examples of failed detection: in depth and thermal data due to the existing obstacle, in depth data due to the face mask occurrence, in RGB model with a projected face on the computer screen.

worst performance with an AP of 97.91%. However, the minor variations between these results make all three techniques applicable in real-world environments similar to the conditions from *easy* image set. The precision-recall curve is visible in Fig. 3(1), showing more details about the performance.

In the medium-difficulty scenario, a similar tendency in performance is observed. Again, the top result is achieved by the RGB approach. The thermal and depth image follows with a drop in performance. The depth model obtains 96.55% and the thermal one a result of 94.07%. The results show that even with some obstacles around the face area, correct detection can still be performed. The occurrence of multiple persons in the image is also not a constrain. Detailed comparison of models' performance is presented in a Fig. 3(2) with the precision-recall curve.

In the hardest scenario the precision-recall curve on Fig. 4(1) shows anomaly in RGB values. This difference is due to the false-positive faces occurrence in some images. In that case, the models which are not based on color information outperform RGB predictors. Overall, the RGB model leads again with a 92.95% result of AP. The thermal model follows with 83.90%. The depth algorithm achieves the worst result with 77.04%. To see the impact of false-positive faces, the performance study is also executed without them. All of the images with projected faces are skipped. The result is a significant improvement of RGB predictor with AP of 99.96%, and the detailed curves of precision and recall are presented in Fig. 4(2) where the anomalies from a Fig. 4(1) are gone.

From our analysis of detection, a tendency to failure of none-RGB based approaches is observed more often. This happens when faces are partially hidden by external obstacles like in a Fig. 5(1). Images with face masks are challenging for depth model which tends to fail with detection like in Fig. 5(2). Except that, performance of thermal images is restricted by low resolution of images in comparison to other cameras. The small resolution affects range of use. In depth images the main constraint likewise in thermal camera is the range of used sensor. In field of acquisition distance, RGB cameras outperforms mentioned modalities. The downside of RGB models is in poor ability of rejecting false face candidates like in Fig. 5(3). Such models can be fooled by mirrors or face images. The 3D based models and thermal detectors are capable of correct predictions in such situations. In general, a major limitation in field of depth and thermal detection is in publicly available data for training purposes. Because of this, our models are trained on notably smaller amount of samples then existing ones for RGB detection. It is very likely that with increase of learning samples the performance will improve and non-RGB models will perform even closer to the RGB images.

Direct positioning of outcomes from this experiment is not possible due to the lack of available benchmarks in the field of depth and thermal face detection. In some studies like [11][13][25] algorithms are evaluated on samples which are captured in laboratory conditions and are far from real-world environment. Faces appear one at a time in the central part of the image. Also, there is a difference in metrics used among publications, where average precision is not calculated in all of them like in [11][25].

In the field of depth data, the first referenced work[11] has the accuracy of correct prediction equal to 93.62%. Similar performance is observed in [25] where depending on light conditions, accuracy varies between 90% and 97%. Chen et al.[6] achieves accuracy of 90% with their descriptor. In the deep learning framework[2] authors present 88.3% value of true-positive and 7.7% false positive detection. Nevertheless, the dataset from this experiment is not available in public anymore.

In the thermal field, for the study of Silva et al.[23] the YOLO AP_{50} metric of 99.75% and AP 78.25% is obtained. However, it is limited by a range of sensors and a single occurrence of faces in samples. In the article from [13] the best model results in a precision of 100% and recall of 98%, but the images are already cropped and do not look challenging for the object detection task. The work of Vukovic et al.[24] appears to be the most similar to ours. They are collecting their test data with multiple faces. Their best result is the precision of 94.33% and the recall of 90.59%.

Considering referenced work and results, the solution presented in this article shows high robustness. A decrease of results appears only in the case of the *hard* dataset. Nevertheless, related works are tested on datasets with the environment in the scenes, which is closer to the laboratory conditions, like presented in this article *easy* dataset. The work described in this paper shows, that applying CNN for depth or thermal face detection is very effective. The introduced dataset creates new comparison possibilities of different techniques in this field.

5 CONCLUSION

In this study, face detection with three different image sensors: RGB, depth, and thermal are explored. Orbbec Astra with Flir Lepton 3.5 cameras are used to collect a new dataset named TMFD Dataset containing RGB, depth, and thermal images, which are captured from the precisely same perspective. Obtained samples include static and non-static scenes with occurrences of single and multiple persons in a scene. All samples are split between easy, medium, and hard scenarios for the detection task. Depending on difficulty, persons are captured in front of different backgrounds, or sensors are placed in a diverse range. A part of samples includes persons wearing accessories like scarfs, hats, glasses, headphones, or face masks to increase the complexity and variety in scenes. In contrast, the referenced publications are evaluated in closer conditions to the laboratory environment with only single face appearances. To perform the detection task, a state-of-art CNN is exploited. Because of the lack of appropriate existing datasets, our models are trained on datasets prepared by us from publicly available images. The pre-trained RGB model outscores other solutions in all three image subsets with AP of 99.99% for *easy* and *medium*. In the *hard* its result is 92.95%. The thermal model performed with AP of 98.98% for *easy*, 94.07 for *medium* and 83.98 for *hard*. The depth one has placed behind them with 97.91%, 96.55%, and 77.04%, respectively. It is impossible to closely compare and place these results among other works due to the lack of publicly available benchmarks. As it was expected, two models do not achieve the average precision of a pre-trained RGB detector, but the scores are high enough to be used in practice. Strengths of non-RGB-based detection have been visible in low-light environments and in rejecting false candidates from mirrors or posters which are rather flat or do not vary in temperature or in low-light environments. This research proves that face detection can be performed with a depth or thermal camera, and an RGB sensor might not be necessary. However, there are some circumstances where these cameras are capable of this task. Their range and low resolution restrict their functionality. In conditions where the sensor is close to the targets, the detection is highly robust. This might simplify the devices used in practice or lower their costs by reducing the number of used cameras. The removal of RGB cameras can also benefit in some cases where personal privacy protection is important.

ACKNOWLEDGMENTS

This work was partly supported by VisuAAL ITN H2020 (grant agreement No. 861091) and the Austrian Research Promotion Agency (FFG) under the Grant Agreement No. 879744.

REFERENCES

[1] Timur Bagautdinov, Francois Fleuret, and Pascal Fua. 2015. Probability occupancy maps for occluded depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2829–2837.

[2] Diego Ballotta, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. 2018. Fully convolutional network for head detection with depth images. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 752–757.

[3] Diego Ballotta, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. 2018. Head Detection with Depth Images in the Wild. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP, Funchal, Madeira, Portugal, January 27-29, 2018.* 56–63.

[4] Tanvi Banerjee, Marilyn Rantz, Mengyuan Li, Mihail Popescu, Erik Stone, Marjorie Skubic, and Susan Scott. 2012. Monitoring hospital rooms for safety using depth images. *AI for Gerontechnology, Arlington, Virginia, US* (2012).

[5] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. 2017. Poseidon: Face-from-depth for driver pose estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5494–5503.

[6] Siyuan Chen, Francois Bremond, Hung Nguyen, and Hugues Thomas. 2016. Exploring depth information for head detection with depth images. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 228–234.

[7] Yuen Kiat Cheong, Vooi Voon Yap, and Humaira Nisar. 2014. A novel face detection algorithm using thermal imaging. In *2014 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*. IEEE, 208–213.

[8] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. 2013. Random Forests for Real Time 3D Face Analysis. *Int. J. Comput. Vision* 101, 3 (February 2013), 437–458.

[9] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2021. Sample and Computation Redistribution for Efficient Face Detection. arXiv:2105.04714 [cs.CV]

[10] Akif Hacinecipoglu, Erhan Ilhan Konukseven, and Ahmet Bugra Koku. 2020. Fast head detection in arbitrary poses using depth information. *Sensor Review* (2020).

[11] RI Hg, Petr Jasek, Clement Rofidal, Kamal Nasrollahi, Thomas B Moeslund, and Gabrielle Tranchet. 2012. An rgb-d database using microsoft's kinect for windows for face detection. In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*. IEEE, 42–46.

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[13] Marcin Kopaczka, Jan Nestler, and Dorit Merhof. 2017. Face detection in thermal infrared images: A comparison of algorithm-and machine-learning-based approaches. In *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 518–529.

[14] Tiemeng Li, Wenjun Hou, Fei Lyu, Yu Lei, and Chen Xiao. 2016. Face detection based on depth information using HOG-LBP. In *2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*. IEEE, 779–784.

[15] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. 2020. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *arXiv preprint arXiv:2006.04388* (2020).

[16] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8759–8768.

[17] Gregory P Meyer, Steven Alfano, and Mink N Do. 2016. Improving face detection with depth. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1288–1292.

[18] Shervin Minaee, Ping Luo, Zhe Lin, and Kevin Bowyer. 2021. Going Deeper Into Face Detection: A Survey. *arXiv preprint arXiv:2103.14983* (2021).

[19] Loris Nanni, Sheryl Brahnam, and Alessandra Lumini. 2019. Face Detection Ensemble with Methods Using Depth Information to Filter False Positives. *Sensors* 19, 23 (2019), 5242.

[20] MI Ouloul, Z Moutakki, K Afdel, and A Amghar. 2015. An efficient face recognition using SIFT descriptor in RGB-D images. *International Journal of Electrical and Computer Engineering* 5, 6 (2015).

[21] Cristina Palmero, Albert Clapés, Chris Bahnsen, Andreas Møgelmose, Thomas B Moeslund, and Sergio Escalera. 2016. Multi-modal rgb–depth–thermal human body segmentation. *International Journal of Computer Vision* 118, 2 (2016), 217–239.

[22] Karen Panetta, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, Holly A Taylor, Arash Samani, et al. 2018. A comprehensive database for benchmarking imaging systems. *IEEE transactions on pattern analysis and machine intelligence* 42, 3 (2018), 509–520.

[23] Gustavo Silva, Rui Monteiro, André Ferreira, Pedro Carvalho, and Luís Corte-Real. 2019. Face Detection in Thermal Images with YOLOv3. In *International Symposium on Visual Computing*. Springer, 89–99.

[24] Tijana Vuković, Ranko Petrović, Miloš Pavlović, and Srđan Stanković. 2019. Thermal Image Degradation Influence on R-CNN Face Detection Performance. In *2019 27th Telecommunications Forum (TELFOR)*. IEEE, 1–4.

[25] Xu Yan and Aoshuang Dong. 2019. Research and Application of Face Detection Algorithm Based on Depth Map. In *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*. IEEE, 258–262.

[26] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A Face Detection Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12993–13000.