2nd German-West African Conference on Sustainable, Renewable Energy Systems SusRES – Kara 2021
Intelligent Automation & Robotic
https://doi.org/10.52825/thwildauensp.v1i.34

Published: 15 June 2021

Concept towards Segmenting Arm Areas for Robot-Based Dermatological In Vivo Measurements

Mateusz Szymanski[1] [https://orcid.org/0000-0002-5991-9906],
Ron van de Sand[1] [https://orcid.org/0000-0002-8975-4030], Esther Tauscher[1], Olaf Rieckmann[2], and Alexander Stolpmann[1][https://orcid.org/0000-0002-3021-2456]

[1]Technical University of Applied Sciences Wildau, Germany

[2]Beiersdorf AG, Germany

Abstract. Dermatological in vivo measurements are used for various purposes, e.g. health care, development and testing of skin care products or claim support in marketing. Especially for the last two purposes, in vivo measurements are extensive due to the quantity and repeatability of the measurement series. Furthermore, they are performed manually and therefore represent a nonnegligible time and cost factor. A solution to this is the implementation of collaborative robotics for the measurement execution. Due to various body shapes and surface conditions, common static control procedures are not applicable. To solve this problem, spatial information obtained from a stereoscopic camera can be integrated into the robot control process. However, the designated measurement area has to be detected and the spatial information processed. Therefore the authors propose a concept towards segmenting arm areas through a CNN-based object detector and their further processing to perform robot-based in vivo measurements. The paper gives an overview of the utilization of RGB-D images in 2D object detectors and describes the selection of a suitable model for the application. Furthermore the creation, annotation and augmentation of a custom dataset is presented.

Keywords: Object Detection, Convolutional Neural Networks, RGB-D images

Introduction

Efficacy and safety studies play a key-role for the development of skin care products. Their core concept is based on performing dermatological laboratory tests on predefined skin areas that are reclusively conducted by specifically trained lab personal. However, the quantity and repeatability of these series of measurements constitutes a major challenge, as they are mostly carried out manually and therefore represent a nonnegligible cost and time factor. A solution to this is the implementation of collaborative robotics for the measurement execution. It goes without saying that common control procedures based on statically programmed software modules cannot be used, since robot-controlled skin measurements must contain additional information on the various body shapes and surface conditions. Furthermore, the position and orientation of the measurement area must be taken into account, which have to be dynamically updated during the measurement process. To solve this problem, information obtained from an image processing system can be integrated into the robot control process. In our previous work, the application of collaborative robotics for the execution of in vivo measurements was proposed [1] by identifying the measurement areas through markers (as

shown in Figure 1). Furthermore, by using a stereoscopic camera spatial coordinates could be determined by performing multiple linear regression. With the information provided, the target coordinates for the robot movement path could be computed as waypoints. However, this approach requires the manual marking of the measurement areas, which seems infeasible in practice. Therefore, the automatic definition of measurement areas can be of high value, as it reduces time and associated costs and may contribute to a higher repeatability. One solution is to use deep learning object detection for image processing with the aim of detecting certain body parts within the image such as upper arm, forearm or hands. More specifically image segmentation methods to retrieve body shapes and contours from the image, which can be further used to identify the respective measurement area. For this, stereoscopic camera depth information being obtained next to RGB (red-green-blue colour image) for each pixel can be used to extract useful features for the subsequent object detection tasks. Zhou et al. [2] and Xing et al. [3] have already shown that the use of RGB-D images (RGB with depth information) can outperform the RGB baseline models. With two types of state-of-the-art object detectors namely, two-/multi-stage or single-stage detector, there is a trade-off between the faster single-stage and the more accurate multi-stage detector [4]. For the online application in a robot control sequence, both speed and accuracy are critical. More recent developments show that this compromise is no longer valid. For example, Wang et al. [5] presented a single-stage focal loss based RetinaNet detector that outperformed the multi-stage detectors in accuracy. Li et al. [6] presented a multi-stage light-head Region-based Convolutional Neural Network, which outperformed the single-stage detectors You Only Look Once (YOLO) and Single Shot Multibox Detector (SSD) in speed and accuracy. Although many researchers dedicated their work to the field of image object detection and many approaches exists throughout the literature [7–11], the combination with robot based in vivo measurements is rarely considered in the past and remains a challenge to date. Therefore, this paper proposes a concept for segmenting arm areas into upper arm, forearm and hand, with the focus on in vivo dermatological measurements performed on the forearm. After this introduction, related works is followed by a concept. Furthermore, this work delivers an overview of state-of-the-art models in terms of applicability to the given application. After selecting a suitable model, it is presented in more detail and the creation of a dataset is described and discussed. Finally, a conclusion is given.

Related Works

Object recognition is one of the most important application fields of image processing [12] and can be used for different objects and tasks, e.g. faces [13], cars [14] or cats and dogs [15]. Histograms of oriented Gradients (HOG) [16], Scale Invariant Feature Transform (SIFT) [17] or Haar-like cascade filters [18] are used for feature extraction and Support Vector Machines (SVM) [19] or Deformable Part-based Models (DPM) [20] are used for classifying based on recognized features. For the detection of more than one object in an image or even objects of different classes it is necessary not only to classify but also to localize the object, this process is called object detection [21]. Most modern object detectors are based on convolutional neural networks (CNN), and although CNNs were proposed by LeCun as early as the 1980s [22] and used for character recognition [23], development stagnated in the early 2000s [21]. With the advancements in GPU computing [24, 25], larger annotated datasets [21] and deeper networks [26], CNNs, like other neural networks, become more advanced and receive more attention in recent years [21].

As mentioned in the previous section, the image identification and segmentation of arm areas is of significant interest to extract the shape and surface of the forearm. A similar problem was coped with in [27] or [28] where body parts for pose estimation where detected. Another model has been proposed in [29] for hand gesture recognition. Other application additionally considered the use of RGB-D data and showed that it can improve the model performance

[30],[31]. Chandra et al. [32] use RGB-D data to segment the limbs, torso and head of a person through a fully convolutional network for the use of a mobility assistance robot. Other works addressed the use of RGB-D images to train CNNs, for example in [33] proposed an R-CNN based model in this context. The depth information was primarily utilized for extracting significant features from the provided image. Another approach is proposed in [34] who implemented a YOLO based approach for detecting objects by using RGB-D images. The authors extended the structure of YOLOv3 by adding another channel in the input layer for the depth information. In addition to the 2D object detection approach, there are 3D object detectors [35, 36] for three-dimensional bounding box estimation. These utilize point clouds and are applied for pose estimation and scene understanding.

Concept

As mentioned in the introduction, the authors presented in their previous work an image processing system for performing robot-assisted dermatological in vivo measurements. As shown in Figure 1, measurement areas were defined using markings on a forearm, which were then used to transform the relevant depth information with the intrinsic and extrinsic parameters into the coordinate system of the camera. With this spatial information and the pixel coordinates of the target positions from the RGB image, target coordinates were computed and passed to the robot control system as waypoints.

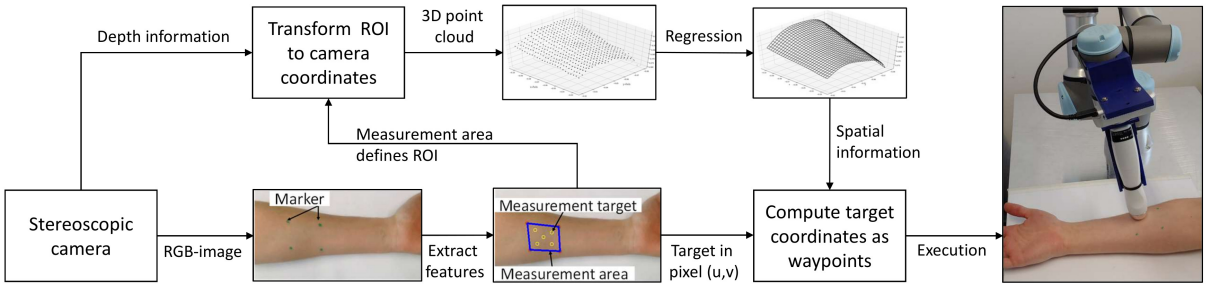

Figure 1. Marker based approach for defining the measurement area [1].

However, placing new markers on the forearm for each measurement seems to be impractical and may not be user-friendly. Therefore, in this work, a concept is presented to replace the markers with an CNN object detector for recognition and localization of the forearm. As input, the object detector receives the depth information in addition to the RGB values for each pixel. The output is expected to be the localization as well as the correct classification of the object contained in the respective picture. For this purpose a bounding box is output for each object, consisting of an anchor point in pixel coordinates and the height and width of the box. Although the authors focus on measurements on the forearm, the neighbouring body parts, hand and upper arm, are detected as well in order to identify the measurement area enclosed by. Due to the functionality of the bounding box determination via the parameters of the Intersection over Union (IoU) [37], the authors expect a higher robustness and better accuracy of the assignment, in particular of the transition areas between hand and forearm as well as forearm and upper arm. Subsequently, the bounding box of the forearm is used as Region of Interest (ROI) to perform an instance segmentation. This is described as determination of the affiliation of pixels to a certain object. For this purpose, the depth information is used to perform a background subtraction as described for different methods in [38] or [39]. Due to the setup and positioning of the camera, the authors are confident that the background is well described in both, the RGB image and especially in the depth information and that the subtraction can be performed reliably. The concept is shown in Figure 2 and the area relevant for this work is highlighted. To give further context on how the segmented information is used in the system, the next steps are also shown.

The pixels segmented to the forearm are used as ROI to transform them into camera coordinates. Unlike the marker based approach, the spatial information of the whole forearm is extracted by a multiple linear regression. With this information and a determined pose, measurement area(s) of a fixed size are fitted on the forearm. From here, the target coordinates are computed and passed to the robot as already shown in Figure 1.

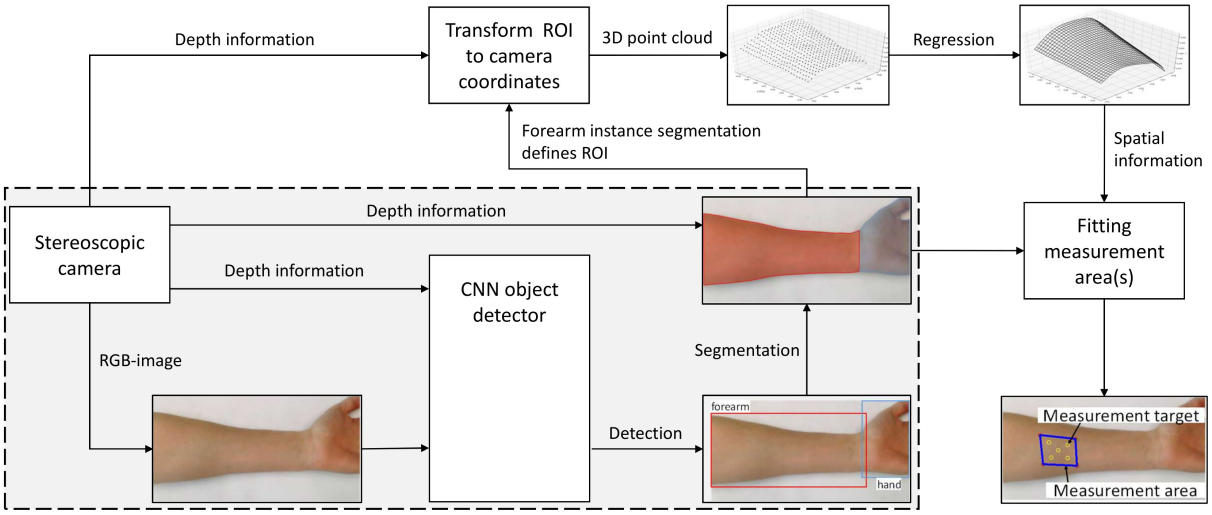

Figure 2. Proposed concept for segmenting arm areas and its further context in the image processing system.

Model Comparison and Selection

For the CNN object detector in the proposed concept, a suitable model must be selected. For this purpose, first an overview of the underlying models and their further developments for the utilization of depth information is given. Afterwards the models are compared and one is selected, which is then described in more detail.

A Brief Overview of Object Detection

In this section, pre-selected object detectors are described and the deployment of depth information based on the RGB models is presented. Selected are Region-based CNN, Fast Region-based CNN, Faster Region-based CNN, You Only Look Once and Single Shot MultiBox Detector.

Region-Based Convolutional Neural Network

Girshick et al. [7] propose a region-based CNN (R-CNN) for object detection. The localization problem is solved by proposing regions for an input image. At test time, about 2000 regions are proposed and each of them is scaled to a defined size. For the region proposal the selective search algorithm [40] is used. The size of the warped region represents the size of the input layer of the CNN. For each region, a feature vector is extracted and classified by a class-specific linear SVM. Gupta et al. [33] use depth images to encode geometric features, namely, horizontal disparity, height above ground and angle (HHA), and generate colour images from them. The authors used two CNNs to extract the features from the RGB images and the HHA features.

Fast Region-Based Convolutional Neural Network

Girshick et al. [8] propose a further development of R-CNN, the Fast R-CNN. The authors address the long training and test time and achieve a higher accuracy compared to R-CNN. First, a feature map is created for a whole image by several convolutional and pooling layers. Then, a ROI is generated for each proposed object and a feature vector is extracted through a

pooling and two fully-connected layers. Via further fully-connected layer the architecture ends in two different output layers. On the one hand a softmax layer for classification of the object and on the other hand an output layer for the bounding box regression. Yuanzhouhan et al. [41] propose RGB-D Fast R-CNN, the network inputs the entire RGB and depth image in two parallel Fast R-CNNs. Two convolutional feature maps are output, which are then concatenated for classification and bounding box regression.

Faster Region-Based Convolutional Neural Network

With the improvements of Fast R-CNN or SPP-net [42], the limitations of modern object detectors are in the region proposal methods. Ren et al. [9] propose Faster R-CNN with a Region Proposal Network (RPN) to address this problem. RPNs basically consist of multiple convolutional layers. Regions are proposed from the convolutional feature maps while simultaneously determining the region boundaries of the object. In principle, Faster R-CNN consists of an RPN for region proposal and Fast R-CNN for object detection. By proposing to share the convolutional features between the RPN and the object detection network, the computational cost is compensated. Ren et al. [43] propose Parallel RCNN for human detection. Parallel RCNN enhances the performance of simultaneously extracting features in RGB and depth images from two CNNs. Further the authors are experimenting with different methods of encoding depth images, e.g. HHA [33].

You Only Look Once

Redmon et al. [10] propose You Only Look Once (YOLO), a new method for object detection. Unlike other object detectors, YOLO views detection not as a region proposal problem but instead as a single regression problem. It processes entire images as input, dividing them into grid cells for this purpose. Each of these cells or cell combinations predicts a certain number of bounding boxes with confidence score. This score indicates how securely an object is located in the box and how well this box fits the contained object. Furthermore, for each box a probability of association with a specific class is predicted. These two are then combined to a final detection. Takahashi et al. [34] propose Expandable YOLO (E-YOLO), in which another channel is added in the input layer for depth information. Further, the authors introduce the 3D IoU with Volume of Overlap and Volume of Union for a better 3D bounding box proposal. Ophoff et al. [44] propose a different approach. They use a separate network stream for the RGB and depth information each and fuse them by a concatenation layer. For the fusion, the authors experiment with different positions of the fusion layer in the network. The authors refer to the proposed model as RGB-D Fusion YOLO.

Single Shot MultiBox Detector

Liu et al. [11] propose with Single Shot MultiBox Detector (SSD) a similar single-stage approach to [10]. SSD takes a whole image as input with ground truth boxes for training. With an input size of 300 x 300 pixel, SSD is faster and more accurate than YOLO with 448 x 448 pixel. The authors use VGG-16 as the base network and then add multiple feature layers. By using multi-scale convolutional bounding boxes that are output to these added feature layers, the bounding box regression problem is solved more efficiently. Further, the authors experiment with different datasets and larger input sizes of images. Sharma and Valles [45] present a base network for processing RGB-D images with SSD. The authors use two parallel streams with convolutional layers for the RGB and depth image. These are then fused using fully connected layers and additional feature layers are added for the detection. This is referred to as RGB-D Fusion SSD.

Model selection

In this work, the model selection is divided in two consecutive steps. First, the RGB baseline is compared in its performance and then the RGB-D models are compared in relation to the baseline, while at the same time evaluating whether the model is suitable for the application

presented. Some researchers may argue that the ranking assignment is not of high scientific value, which is indeed the case, but the wide variety of hardware and architectures and datasets makes it difficult to compare objectively. This applies in particular to the assessment of suitability, which is why it has to be regarded as subjective and will subsequently be addressed in the conclusion.

For the comparison of baseline performance, the metrics of mean Average Precision (mAP) and test time are used. In addition, the used architecture and dataset is given. The mAP gives the average over the precision over all classes of a dataset. The test time indicates how long the model needs to process one image during runtime. The comparison is given in Table 1.

Table 1. Comparison of different object detector models.

Model	Architecture	Dataset	mAP	Test time [s]
R-CNN [7]	AlexNet	VOC07	58.5	9.8
	VGG16	VOC07	66.0	47.0
Fast R-CNN [8]	VGG16	VOC07	66.9	0.32
Faster R-CNN [9]	VGG16	VOC07+12	73.2	0.2
YOLO [10]	VGG16	VOV07+12	66.4	0.048
	GoogLeNet	VOC07+12	63.4	0.022
SSD [11]	VGG16	VOC07	74.3	0.017

Table 1 shows that single-stage detectors such as YOLO or SSD can achieve the accuracy of Faster RCNN and are thereby significantly faster in test time. The next step is to compare the RGB-D models from the previous section. The models are benchmarked against their own, if known, RGB baseline to show the effect of using additional depth information. Since test time is often not specified, the effect on it is further estimated with Δ for better, $\circ$ for neutral and ∇ for worse compared to the baseline. The same metric is finally used to assess whether the model could be used for use case presented in this paper. The results are shown in Table 2.

Table 2. Comparison and rating of different RGB-D models.

Model	Dataset	mAP	Test time [s]	Suitable?
R-CNN	NYUD2	19.7	-*	
R-CNN + HHA [33]	NYUD2	32.5	-* ∇	∇
Fast R-CNN	B3DO	39.9	-*	
RGB-D Fast R-CNN [41]	B3DO	41.9	-* ∇	∇
Faster R-CNN	custom	90.0	-*	
Parallel R-CNN [43]	custom	91.5	-* ∇	$\circ$
E-YOLO [34]	custom	-*	0.023	Δ
YOLOv2	KITTI	39.87	-*	
RGB-D Fusion YOLOv2 [44]	KITTI	48.16	-* ∇	Δ
RGB-D Fusion SSD [45]	Princeton + Washington RGB-D	99.43	-*	Δ

* not available

The additional processing of depth information can basically be divided into three methods. First method, the encoding of the depth information into a three-channel colour image and processing in a parallel CNN. Similarly, in the second method the depth information is processed in a parallel CNN stream as a single channel input and then fused, with different approaches to when the fusion takes place in the network. And finally for the third method adding another input channel for the depth information. In general, the processing of additional depth information is considered to have a negative impact on the test time, since it requires

additional computations. The authors consider a test time of about 0.033 to 0.05 seconds per image, which corresponds to 20 to 30 frames per second, to be suitable. Therefore, the models based on the R-CNN family must be classified as not suitable for our application. Choosing between E-YOLO, RGB-D Fusion on YOLO or SSD, the authors opt for the latter. The decision is based on four considerations:

- First, SSD performs better in terms of accuracy and test time on the baseline (as shown in Table 1).
- Second, starting from the baseline, there are opportunities to sacrifice speed for accuracy.
- Third, the method of using a separate CNN stream for depth information is well known, as shown by other methods compared here.
- And fourth, the application for which the model is proposed, the grasping of different objects with a mobile robot arm, has certain similarity to our intention.

Model Description

With the selection of the RGB-D Fusion SSD model, it is described in more detail. The RGB and depth image are scaled to 200 x 200 pixel images and processed in two parallel CNN streams for convolution and feature extraction. These consist of five convolutional layers with maxPooling after the second, fourth and fifth layer. In addition, batch normalization is performed in the depth CNN stream after each convolutional layer. After feature extraction, the two CNN streams are merged via a concatenation through three fully connected layers [45]. From here on, the original SSD architecture [11] is used with four additional convolutional layers ending in a softmax layer for classification followed by a non-maximum suppression layer for fitting the bounding box. The whole architecture is shown in Figure 3.

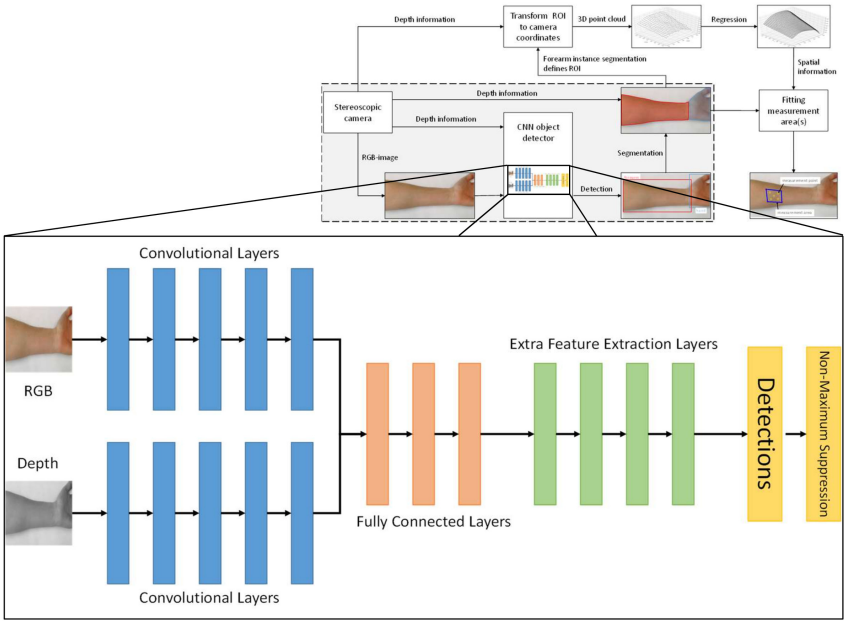

Figure 3. Network Overview of SSD with RGB-D Architecture derived from [45].

With the presentation of a concept and selection of a model, the following section deals with the creation of a dataset.

Dataset

For the very specific use case of detecting forearms, a custom dataset is created. The creation consists of three steps, data acquisition, annotation, and augmentation of the data.

For the data acquisition the setup from previous work is used, consisting of the collaborative robot UR3e from Universal Robots and an Intel RealSense D435i stereoscopic camera, which is attached on top of the robot's arm. The robot moves to five different positions in one pass. Each of these positions is randomized, within certain limits, to allow variation in depth and viewing angle. At each position, an RGB and depth image is captured. For each forearm, 2 to 3 passes are made in different poses. A total of 750 images are taken from 35 people. Further on, the images have to be labelled manually with the areas corresponding to the forearm, hand and upper arm. Each of them represents an individual label which consists of selected pixel coordinates. Due to the setup described above, there are three different possible combinations. Figure 4 shows an example of the annotation process for three images, with forearm and upper arm in row (a), forearm and hand in row (b) and all three labels in row (c).

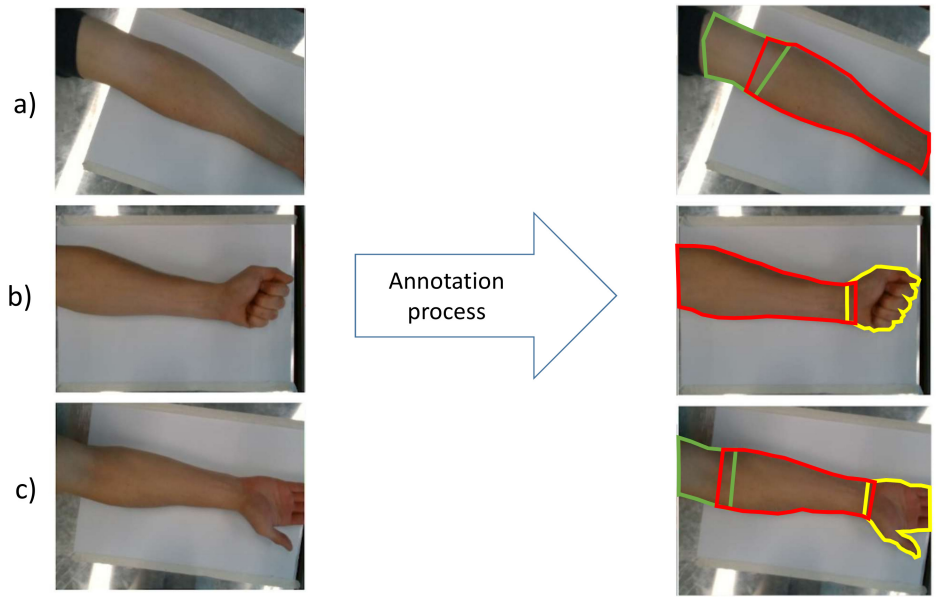

Figure 4. Annotation process for (a) forearm and hand, (b) forearm and upper arm and (c) all three labels.

In the final step, the raw images are augmented to a larger database. In order to achieve a certain robustness against changes in brightness and illumination, the image is converted to the HSV colour space and augmented in the H-S, H-V, S-V and HSV channels. The change in the corresponding channels is randomized within certain limits. Afterwards it is transferred back to the RGB colour space. To provide greater variation, rotations of 90°, 180° or 270° are applied. This results in seven different modes of augmentation. Attention must be given during rotation in regard to the annotated coordinates; they must also be rotated accordingly. Figure 5 shows an example of the augmentation for three images. In total the dataset consists of 6000 annotated images.

Discussion

It should be noted that model selection is not a trivial task and developers may draw different conclusions, as quantitative comparison of all models is only possible to a limited extend. Therefore, future work will address this issue in more detail by comparison multiple architectures in terms of accuracy and computational resources. Further, other methods to use depth information in 2D object detectors will be considered. Especially E-YOLO is interesting because of its simplicity and lightweight implementation. Moreover, the single channel input of the depth stream can be replaced by a three channel one, which allows the use of encoded depth images. Another field of research intersecting with this application are 3D object detectors. As aforementioned, in this work, the transformation of the camera depth

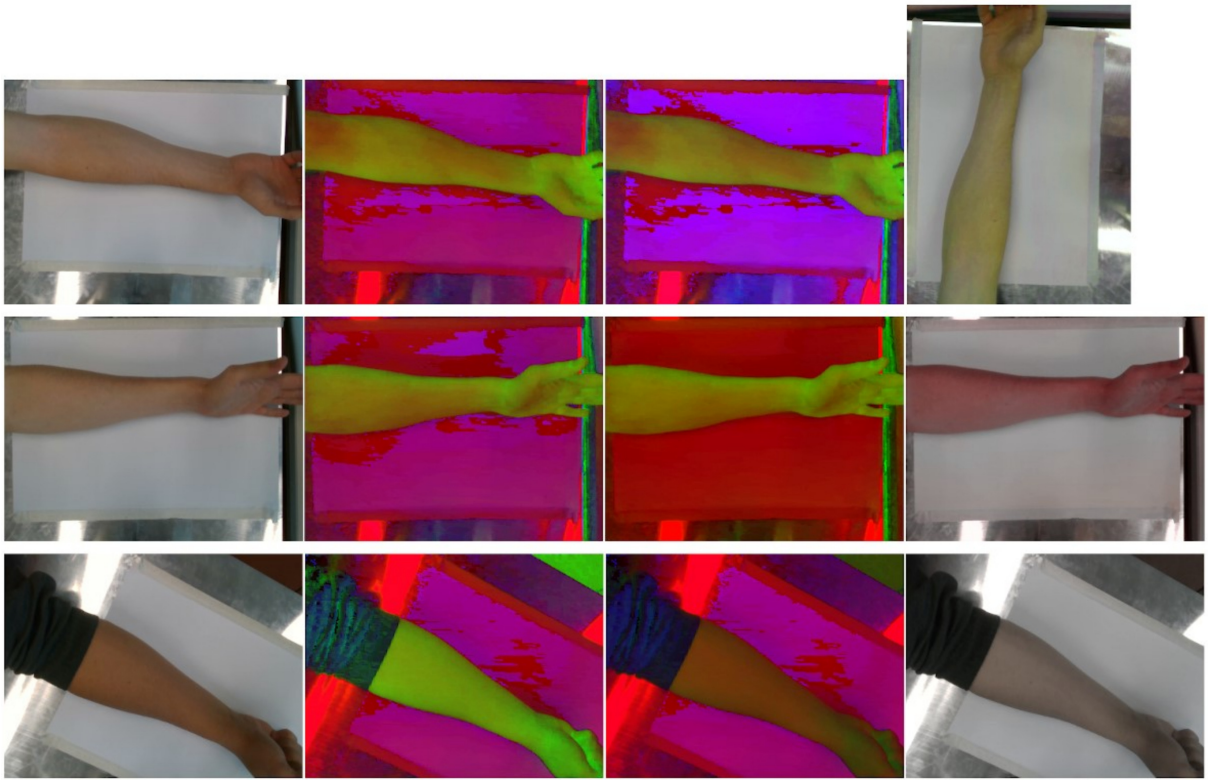

Figure 5. Image in RGB, image in HSV, augmented image in HSV, augmented image in RGB (from left to right), first row: H-V modified and rotated by 270°, second row: S-V modified, third row: H-S modified.

data into the 3D space is conducted after detecting the object of interest. Thus, only the area of interest is transformed while comparatively few computational resources are occupied. Despite that, the application of 3D object detectors could also be favourable. However, this may include the transformation of all depth points instead of a designated area only, which can be computationally expensive. Finally, further research projects will also focus on the generalisation ability of the model to generalise well by enhancing the dataset with, for example, different genders and skin colours.

Conclusion

In this paper, a concept for the detection and segmentation of arm areas using CNN-based object detectors with the context of automated execution of robot-based dermatological in vivo measurements is presented. The concept is based on previous work and gives further details on the use of the segmented arm areas in the image processing system. By deploying a stereoscopic camera both RGB images as well as associated depth information becomes available, which is considered in the derived concept. Furthermore, this paper gives an overview about methodologies to apply depth information in 2D object detectors and points out challenges within the described application framework. Besides, the creation, annotation and augmentation of a custom dataset is presented.

References

[1] M. Szymanski, R. van de Sand, O. Rieckmann, and A. Stolpmann, “Robotergestützte dermatologische in-vivo-Messungen,” *atp magazin*, vol. 62, no. 11-12, pp. 78–85, 2020.

[2] K. Zhou, A. Paiement, and M. Mirmehdi, “Detecting humans in RGB-D data with CNNs,” in *Proceedings of the fifteenth IAPR International Conference on Machine Vision Applications*, Piscataway, NJ: IEEE, 2017, pp. 306–309.

[3] Y. Xing, J. Wang, X. Chen, and G. Zeng, “2.5D convolution for RGB-D semantic segmentation,” in *2019 IEEE International Conference on Image Processing*, Piscataway, NJ: IEEE, 2019, pp. 1410–1414.

[4] P. Soviany and R. T. Ionescu, “Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction,” in *SYNASC 2018*, [Los Alamitos, Calif.]: IEEE Computer Society, 2018? Pp. 209–214.

[5] X. Wang, P. Cheng, X. Liu, and B. Uzochukwu, “Focal loss dense detector for vehicle surveillance,” in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV2018)*, Piscataway, NJ: IEEE, 2018, pp. 1–5.

[6] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, *Light-head R-CNN: In defense of two-stage object detector*.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 580–587.

[8] R. Girshick, “Fast R-CNN,” in *2015 IEEE International Conference on Computer Vision*, Piscataway, NJ: IEEE, 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *29th IEEE Conference on Computer Vision and Pattern Recognition*, Piscataway, NJ: IEEE, 2016, pp. 779–788.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905, Cham: Springer International Publishing, 2016, pp. 21–37.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[13] J. Meng, Y. Gao, X. Wang, T. Lin, and J. Zhang, “Face recognition based on local binary patterns with threshold,” in *IEEE International Conference on Granular Computing (GrC), 2010*, X. Hu, Ed., Piscataway, NJ: IEEE, 2010, pp. 352–356.

[14] T. Moranduzzo and F. Melgani, “Detecting cars in UAV images with a catalog-based approach,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6356–6367, 2014.

[15] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, “Cats and dogs,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*, Piscataway, NJ: IEEE, 2012, pp. 3498–3505.

[16] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR 2005*, C. Schmid, C. Tomasi, and S. Soatto, Eds., Los Alamitos, Calif: IEEE Computer Society, 2005, pp. 886–893.

[17] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *CVPR 2001*, Los Alamitos, Calif: IEEE Computer Society, 2001, pp. I-511-I–518.

[19] O. Chapelle, P. Haffner, and V. N. Vapnik, “Support vector machines for histogram-based image classification,” *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 1055–1064, 1999.

[20] R. Girshick, F. Iandola, T. Darrell, and J. Malik, “Deformable part models are convolutional neural networks,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Piscataway, NJ: IEEE, 2015, pp. 437–446.

[21] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, *Object detection with deep learning: A review*.

[22] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, *A survey of the recent architectures of deep convolutional neural networks*, 2020.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] D. Steinkraus, I. Buck, and P. Y. Simard, “Using GPUs for machine learning algorithms,” in *Proceedings / Eighth International Conference on Document Analysis and Recognition, 2005*, Los Alamitos, Calif.: IEEE Computer Society, 2005, 1115–1120 Vol. 2.

[25] K.-S. Oh and K. Jung, “GPU implementation of neural networks,” *Pattern Recognition*, vol. 37, no. 6, pp. 1311–1314, 2004.

[26] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*.

[27] Q. Dai, J. Qiao, F. Liu, X. Shi, and H. Yang, “A human body part segmentation method based on markov random field,” in *International Conference on Control Engineering and Communication Technology (ICCECT), 2012*, Piscataway, NJ: IEEE, 2012, pp. 149–152.

[28] A. Jalal, A. Nadeem, and S. Bobasu, “Human body parts estimation and detection for physical sports movements,” in *2019 2nd International Conference on Communication, Computing and Digital Systems (C-CODE)*, Piscataway, NJ: IEEE, 2019, pp. 104–109.

[29] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, “Robust part-based hand gesture recognition using kinect sensor,” *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[30] C. Plagemann, V. Ganapathi, D. Koller, and S. Thrun, “Real-time identification and localization of body parts from depth images,” in *IEEE International Conference on Robotics and Automation (ICRA), 2010*, Piscataway, NJ: IEEE, 2010, pp. 3108–3113.

[31] N. Mohsin and S. Payandeh, “Localization and identification of body extremities based on data from multiple depth sensors,” in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Piscataway, NJ: IEEE, 2017, pp. 2736–2741.

[32] S. Chandra, S. Tsogkas, and I. Kokkinos, “Accurate human-limb segmentation in RGB-D images for intelligent mobility assistance robots,” in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, IEEE, 7.12.2015 - 13.12.2015, pp. 436–442.

[33] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, “Learning rich features from RGB-D images for object detection and segmentation,” in *Computer vision - ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, Ed., vol. 8695, Cham: Springer, 2014, pp. 345–360.

[34] M. Takahashi, Y. Ji, K. Umeda, and A. Moro, “Expandable YOLO: 3D object detection from RGB-D images,” in *”2020 21st International Conference on Research and Education in Mechatronics (REM)”*, IEEE, 2021-01-12, pp. 1–5.

[35] S. Song and J. Xiao, “Deep sliding shapes for amodal 3D object detection in RGB-D images,” in *29th IEEE Conference on Computer Vision and Pattern Recognition*, Piscataway, NJ: IEEE, 2016, pp. 808–816.

[36] D. Xu, D. Anguelov, and A. Jain, “Pointfusion: Deep sensor fusion for 3D bounding box estimation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Piscataway, NJ: IEEE, 2018, pp. 244–253.

[37] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Piscataway, NJ: IEEE, 2019, pp. 658–666.

[38] K. Greff, A. Brandão, S. Krauß, D. Stricker, and E. Clua, “A comparison between background subtraction algorithms using a consumer depth camera,” in *Proceedings of the International Conference on Computer Vision Theory and Applications*, SciTePress - Science and and Technology Publications, 24.02.2012 - 26.02.2012, pp. 431–436.

[39] E. J. Fernandez-Sanchez, J. Diaz, and E. Ros, “Background subtraction based on color and depth using active sensors,” *Sensors (Basel, Switzerland)*, vol. 13, no. 7, pp. 8895–8915, 2013.

[40] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective search for object recognition,” *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[41] C. Yuanzhouhan, S. Chunhua, and T. S. Heng, “Exploiting depth from single monocular images for object detection and semantic segmentation,” *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 26, no. 2, pp. 836–846, 2017.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[43] X. Ren, S. Du, and Y. Zheng, “Parallel RCNN: A deep learning method for people detection using RGB-D images,” in *CISP-BMEI 2017*, Q. Li, Ed., Piscataway, NJ: IEEE, 2017, pp. 1–6.

[44] T. Ophoff, K. van Beeck, and T. Goedemé, “Exploring RGB+depth fusion for real-time object detection,” *Sensors (Basel, Switzerland)*, vol. 19, no. 4, 2019.

[45] P. Sharma and D. Valles, “Backbone neural network design of single shot detector from RGB-D images for object detection,” in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, 10/28/2020 - 10/31/2020, pp. 0112–0117.