Identification and Classification of Human Body Parts for Contactless Screening Systems: An Edge-AI Approach

Diogo Rocha[1], Pedro Rocha[1], Jorge Ribeiro[1], and Sérgio Ivan Lopes[1,2(✉)]

[1] ADiT-LAB, Instituto Politécnico de Viana do Castelo, Rua Escola Industrial e Comercial Nun'Álvares, 4900-347 Viana do Castelo, Portugal
sil@estg.ipvc.pt

[2] IT - Instituto de Telecomunicações, Campus de Santiago, 3810-193 Aveiro, Portugal

Abstract. Continuous monitoring of vital signs like body temperature and cardio-pulmonary rates can be critical in the early prediction and diagnosis of illnesses. Optical-based methods, i.e., RGB cameras and thermal imaging systems, have been used with relative success for performing contactless vital signs monitoring, which is of great value for pandemic scenarios, such as COVID-19. However, to increase the performance of such systems, the precise identification and classification of the human body parts under screening can help to increase accuracy, based on the prior identification of the Regions of Interest (RoIs) of the human body. Recently, in the field of Artificial Intelligence, Machine Learning and Deep Learning techniques have also gained popularity due to the power of Convolutional Neural Networks (CNNs) for object recognition and classification. The main focus of this work is to detect human body parts, in a specific position that is lying on a bed, through RGB and Thermal images. The proposed methodology focuses on the identification and classification of human body parts (head, torso, and arms) from both RGB and Thermal images using a CNN based on an open-source implementation. The method uses a supervised learning model that can run in edge devices, e.g. Raspberry Pi 4, and results have shown that, under normal operating conditions, an accuracy in the detection of the head of 98.97% (98.4% confidence) was achieved for RGB images and 96.70% (95.18% confidence) for thermal images. Moreover, the overall performance of the thermal model was lower when compared with the RGB model.

Keywords: Edge-AI · Identification · Classification · Validation · Thermal · RGB

S. Paiva et al. (Eds.): SmartCity360° 2021, LNICST 442, pp. 92–103, 2022.
https://doi.org/10.1007/978-3-031-06371-8_7

1 Introduction

Due to the recent COVID-19 outbreak, new problems emerged in the way that our health professionals work. COVID-19 is known for its high transmission rate through the air, being its progression typically controlled in three stages [1]: 1) social distancing; 2) quarantine suspects and 3) isolate infected. Statistics point to approximately 80% of infected individuals develop light symptoms as low fever, headache, or myalgia, although there are other 15% that can have the low symptoms and as well pneumonia or breathing problems. The remaining 5%, besides the previous symptoms, are likely to develop skeptic shock, respiratory failure, or multi-organ failure and need additional attention by healthcare professionals and increased contact with the patients, which may be a source of additional infection [2]. Therefore, continuous contactless vital signs screening has a huge potential for reducing the physical contact between patients and healthcare professionals.

This work was developed under the R&D project CoViS[1], whose aim is to design and develop a real-time contactless health monitoring system, that uses a multimodal approach based on Doppler radar techniques [3] to measure the cardiorespiratory rates, and thermography images based on infrared to assess the human body temperature. Therefore, to improve the system performance, the prior identification of the Regions-of-Interest (RoIs)—head, torso, arms—of the individual under screening is of major importance. In that context, with the images captured by the cameras, the software must be capable to identify and classify the individual body parts, lying and stand or even in other perspectives, and perform the screening task properly contextualized, accordingly.

We investigated the use of Convolutional Neural Network (CNN) in an open-source implementation using small electronic devices to the classification of human body parts, in particular when a human is lying for example in a bed. The main focus of this work is to detect parts of the body in a specific position through RGB and Thermal images, and not detect the segmentation or pose estimation of the human body. In this sense, the work is based on the data gathering from a thermal camera and the position of the human body in the context of a healthcare environment (ex. a person lying in bed). To create the prediction model we used a specific camera in different positions to obtain the images (RGB and Thermal) for the dataset. Thus, the previous processing task was analyzing the images and execute the interpolation in order to adapt the size of the images to be used with the CNN implementation. The method uses a supervised learning model that can run in edge devices, e.g. Raspberry Pi 4,

This paper will be organized in sections, starting with Sect. 2 called Related Work, which will showcase projects that use technologies necessary for the future implementation of this project. Section 3 called Adopted Methodology will contain our data set preparation and the AI algorithm selection that we developed

[1] *CoViS—Contactless Vital Signs Monitoring in Nursing Homes using a Multimodal Approach*, Project website: https://covis.wavecom.pt/.

throw this project. Section 4 will be the results of our AI algorithm. Finally, Sect. 5, where the conclusions are taken and the future work is defined.

2 Related Work

Due to the success of the applicability of Artificial Intelligence techniques in computer vision, in the last years have been presented some other approaches using Camera Pose Estimation with Deep Learning [4], in particular for image classification, image segmentation, object detection, and many more image characteristics. Using the basis of Convolutional Neural Networks, different approaches have been presented for example using the idea for regressing the absolute camera pose from an RGB image or thermal/heat image use. Besides the promising results, in general, the resulting accuracy was sub-optimal, compared to classic feature-based solutions in particular in a particular position of the body in the camera field. This led to a surge of learning-based pose estimation methods that can complement the approach of classifying human body parts in a specific position. In the future we intend to explore and complement this work and their context applicability using other Deep Learning Approaches as PoseNet [5], LSTM-Pose, Bayesian and Support Vetor Machine Pose Net [6], and 3D human pose estimation classification and recognition [7–9] with different encoders (GoogleNet, densenet, for example) and with different opensource language/frameworks implementations (tensorflow posenet (1), MediaPipe (2), for example). On another end, in the last years, different opensource implementations have been presented to classify or recognize characteristics in images. In the last years a huge number of works have been exploring CNN implementation, mainly using opensource approaches, for example, the Tensorflow [12], Pytorch [10], Keras [11] or other implementations in Java, C++, and python, to present a few. In this work we followed the exploration of the Tensorflow lite implementation [12,13], by the fact of the promising results for identifying and classify characteristics in images as well as the ability and capacity to run in edge devices, e.g. Raspberry Pi 4. In the future, we intend to explore other implementations or even improve some in order to better accurate the results of the models and other specificities of Convolutional Neural Networks.

Plagemann et al. [14] propose a point detector particularly designed for analyzing the human shape. The interest points, which are based on identifying geodesic extreme points, coincide with salient points of the body, which can be classified as, e.g., hand, foot or head using local shape descriptors. According to the authors, their approach provides a natural way of estimating a 3D orientation vector for a certain interest point, that can be used to simplify the classification problem as well as estimate the orientation of the body parts in space. The training set consists of 789 recorded frames from a different sequence, resulting in 6312 patches extracted at interest point locations. They evaluated the classifier in the test set and it proved to be almost perfect with 98% accuracy for the patches containing the head. The respective numbers for hands and feet were 82% and 79%. This method presents an increase in performance over the state-of-the-art alternatives.

In [15], Romero et al. present a method for estimating 2D human pose from video using only optical flow. Their method, called *FlowCap* method uses a Kalman filter to propagate body part positions and velocities over time and a regression method to predict 2D body pose from part centers. No range sensor is required and FlowCap estimates 2D human pose from monocular video sources containing human motion. Such sources include hand-held phone cameras and archival television videos. The authors, also demonstrate 2D body pose estimation in a range of scenarios and show that the method works with real-time optical flow. The method was trained using the HumanEva training set [16], composed of approximate 7000 training images of the full-body, and the authors generated two generic datasets: The upper body dataset is composed of approximately 7, 000 training examples, while the full-body dataset has approximately 14, 000. The results suggest that the training data generated [15] could be used to directly train a CNN to estimate pose from flow (and image data).

Cao et al. [17], propose to efficiently detect the 2D pose of multiple people in an image. The approach uses a nonparametric representation, which refers to Part Affinity Fields (PAFs), to learn to associate body parts with individuals in the image. The architecture encodes global context, allowing a greedy bottom-up parsing step that maintains high accuracy while achieving real-time performance, irrespective of the number of people in the image. The architecture is designed to jointly learn part locations and their association via two branches of the same sequential prediction process. In this paper, the authors presented an explicit non-parametric representation of the key-points association that encodes both position and orientation of human limbs. Second, they designed an architecture for jointly learning parts detection and parts association. Third, they demonstrate that a greedy parsing algorithm is sufficient to produce high-quality parses of body poses, that maintain efficiency even as the number of people in the image increases. They showed representative failure cases as well.

3 Adopted Methodology

Figure 1 illustrates the adopted methodology, which consists of four distinct steps:

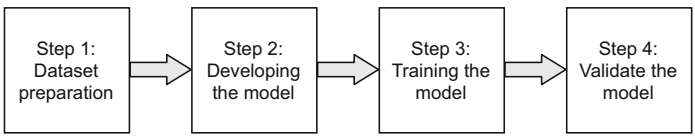

Fig. 1. Adopted methodology.

The first step consisted of the dataset preparation. All the images were obtained with the FLIR E54 [18] camera. The FLIR E54 thermal camera has

two built-in cameras, one based on a CMOS RGB sensor, and the other based on a thermal imaging sensor. The thermal imaging sensor has a resolution of 320×240 pixels and operates in the interval from -20 to $120\,^{\circ}\mathrm{C}$, with a temperature accuracy of $\pm 0.3\,^{\circ}\mathrm{C}$. On the other hand, the RGB chipset has a resolution of 1280×960 pixels. The camera has built-in Wi-Fi connectivity and is also capable of performing live-streaming. The second step focused on developing the model that will enable the identification and the classification of human body parts, more specifically the head, torso, and arms, which has been done by manually evaluating the pictures taken in the previous step. The third step consists of training the model, which also included the model tuning iterations. Lastly, the final step consisted of the model validation with new images representing the practical application case under study.

3.1 Dataset Preparation

The dataset was obtained during three weeks, where several subjects have participated in different poses to simulate the application case under study. Figure 2 depicts the experimental apparatus used to obtain the dataset, where three camera positions have been considered. The images were obtained using the thermal camera FLIR E54, being that we obtained 876 images, 438 thermal and 438 RGB, of which 42 correspond to Setup 1, and have been taken from a low tripod with about 1.2 m in height and a distance from the subject of 1.50 m. Setup 2 was obtained using the tripod at approximately 0.70 m height, which resulted in a distance to the subject of 2.1 m. In Setup 3, 279 pictures have been acquired using a giraffe tripod at 2 m height and a distance to the subject of 2.50 m.

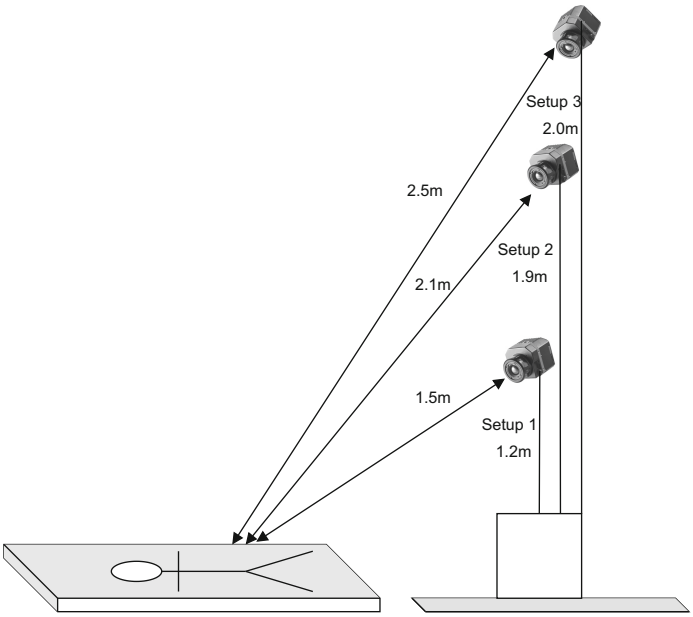

Fig. 2. Image acquisition procedure with the three setups evaluated.

Figure 3 illustrates two example images obtained with the FLIR E54 camera, RGB and Thermal, respectively, with all relevant Regions of Interest (RoI) identified before labeling.

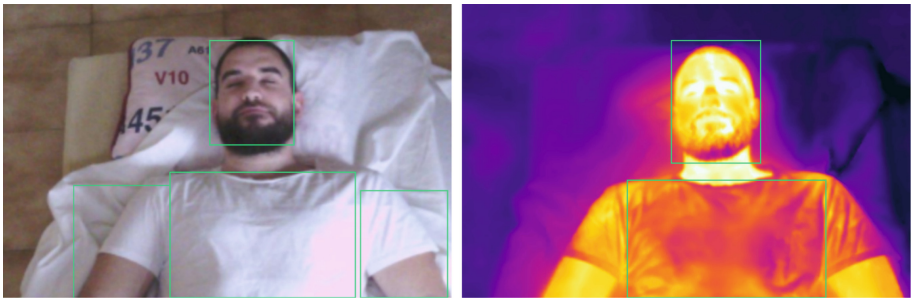

Fig. 3. Regions of Interest (RoI) definition. Example of images (RGB+Thermal) obtained with the FLIR E54 camera.

3.2 Model Definition

Our model was conceptualized firstly with a focus on the head, torso, and arms. However, the detection of the head and torso have been prioritized due to its applicability for contactless vital signs monitoring applications. In different environments, several scenarios have been portrayed for distinct patient's positions at a nursing room bed, in a way to make the experiment as close as possible to a real situation. We used FLIR E54 to capture the RGB and Thermal datasets, which have been captured at the same time instant.

Contactless vital signs monitoring techniques rely on different technologies that need to be optimally calibrated and aligned for increased accuracy. In that sense, the need for detecting different parts of the body—in both RGB and thermal images—is of great value to increase the accuracy of such methods. This way, the proposed model for detection and identification has been prioritized by the relevance of the body parts: 1) head, 2) torso, and 3) arms. The body parts detection will be performed and evaluated using both types of images for all the datasets produced.

All the acquired RGB images that are part of the three produced datasets have been labeled accordingly to the model previously defined. Regarding the thermal images, we opted to ignore the arms part because, in the context of the application we are targeting, only the head and torso are being considered, which simplifies the model and facilitates its deployment in resource-constrained edge devices, since fewer resources will be needed, leading to the improvement of the overall model efficiency.

3.3 Model Training

The two types of images captured (RGB and Thermal) have been then used to spot the differences in the detection and identification of body parts. The training

process was performed using the GPU (Graphical Processing Unit) NVIDIA GeForce GTX 1650 With Max-Q Design, 4 GB GDDR5 and both CUDA and CUDNN extensions have been enabled to speed up processing.

First, the RGB images subset has been trained with 479 images, that have been labeled based on the body parts model previously defined, which included the head, torso, and arms. Then, the SSD MobileNet V2 FPNLite model [13,20,21] was used with a resolution of 640×640 in every image to initialize the train of the RGB model. The train was performed during 2.000 iterations and, in the end, a tune has been performed, we altered the iterations number, which resulted in more than 50.000 iterations. The total duration of the train was approximately 48 h. Secondly, the subset of the thermal images (also 479 images), has been trained and labeled based on the body parts model previously defined, but which included only the head and torso. Then, the SSD MobileNet V2 FPNLite model [13,20,21] was used with a resolution of 320×320 in every image. The train was performed during 1.000 iterations and, in the end, a tune has been performed, we altered the iterations number, which resulted in more than 50.000 iterations. The total duration of the train was approximately 38 h. The technique used for validation was the train/test split.

- After acquiring the images, they were converted from resolution 1280×960 to 640×640 in the RGB dataset, and from the resolution 320×240 to 320×320 for the thermal images. This step has been carried out to achieve optimal performance with the pre-trained CNN implementation. In this case, each image was associated with specific ground-truth labeling and the classification categories (Head, Torso, Arms).
- Then the dataset was split into two subsets, i.e., training and validation sets. The training set consists of 115 images randomly selected from all the dataset, and the validation set consists of 341 images that remained and that makes 456 images. The 456 images in total, trained and validated are still less than the 479 that we total captured, that's because we decided to exclude some images in order to get better results, i.e., lack of light or blurred images. That resulted in excluding 41 RGB images and 23 thermal images, that lead 438 RGB images and 456 thermal images in total.

Typically, during a CNN evaluation some specific metrics based on False Positives, False Negatives, True Positives, True Negatives (presented as the confusion matrix), model accuracy, precision, root-mean-square error, F1-score, as well as some CPU and GPU performance processing metrics, such as the mean Average Precision (mAP) and the Intersection over Union (IoU)—which determines how many objects were detected correctly and how many false positives were generated.

To evaluate the robustness of the MobileNet V2 FPNLite model [13,20,21], real images without ideal conditions from outside our dataset, have also been used. Moreover, a conventional webcam has also been used to evaluate in a live stream the quality of the obtained data. Both MobileNet models were trained and evaluated with an Open Source TensorFlow Object, in particular, the detection

accuracy (DA) and confidence level (CL), described in the next section. Figure 4 depicts the implemented MobileNet V2 SSD architecture.

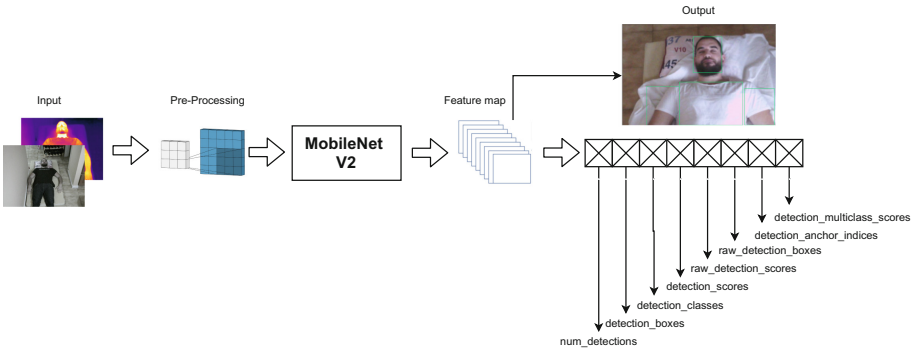

Fig. 4. MobileNet V2 SSD architecture.

4 Results and Discussion

The results obtained with the RGB dataset, varied depending on the setup used, cf. Fig. 2, we evaluated them in detection accuracy (DA) and confidence level (CL) which was computed based on the percentage of cases with confidence above 90%. In general, head detection produced the best results. The torso generated pretty satisfactory results, although not as good as the head detection. Lastly, the arms, which gave us subpar results compared to the other body parts. The results obtained are compiled in Table 1.

Still, in the RGB dataset, the first setup produced an head DA of 95.51% with a CL of 95.23%, and torso DA of 90.50% with a CL of 83.33%. The arms had a DA of 38.87% and CL of 22.62%. These results were obtained from 42 images with the conditions described in Sect. 3.1. On the second setup, the DA of the head is 97.82% with an LoC of 95.76%, with the torso DA being 77.19% and CL of 61.86%. The arms produced a DA of 46.41% and a CL of 19.49%. To obtain these results, 118 images were tested. And in the third and last one, the DA of the head is 99.97% with a CL of 100.00%, the torso DA 94.45% with a CL of 92.45%, and the arms DA 63.27% with a CL of 52.35%. These results refer to 278 images.

With all this, the general results for all the RGB datasets were 98.97% regarding the head DA with a CL of 98.40%, 89.42% regarding the torso DA with a CL of 83.33% and a DA of 56.39% for the arms with a CL of 40.87%.

Regarding the thermal dataset, we had the same three setups, but we opted to not label the arms. With this in mind, the first setup gave us a head DA of 90.69% with a CL of 88.89% and a torso DA of 46.99% with a CL of 37.78%. These results came from 45 images with the conditions described in Sect. 3.1. On the second one, the DA of the head is 97.78% with a CL of 96.88%, with the

Table 1. Results summary for the three evaluated setups.

	Setup 1		Setup 2		Setup 3		Total	
	RGB	Thermal	RGB	Thermal	RGB	Thermal	RGB	Thermal
Accuracy								
Head	95.51%	90.69%	97.82%	97.78%	99.97%	96.93%	98.97%	96.70%
Torso	90.50%	46.99%	77.19%	41.54%	94.45%	42.10%	89.42%	42.64%
Arms	38.87%	–	46.41%	–	63.27%	–	56.39%	–
Confidence								
Head	95.23%	88.89%	95.76%	96.88%	100.00%	95.05%	98.40%	95.18%
Torso	83.33%	37.78%	61.86%	33.59%	92.45%	31.45%	83.33%	32.89%
Arms	22.62%	–	19.49%	–	52.34%	–	40.87%	–
Number of images	42	45	118	128	278	283	438	456

Fig. 5. Example of setup 3 results (left - normal images, right - images with artifacts).

torso DA being 41.54% and CL of 33.59%. To obtain these results, 128 images were analyzed. Finally, in the last one, the DA of the head is 96.93% with a CL of 95.05% and the torso DA 42.10% with a CL of 31.45%. These results came from 283 images. These results gave the thermal model an overall DA for the head of 96.70% with a CL of 95.18% and a torso DA of 42.64% with a CL of 32.89%, with a total of 456 images tested.

To visualize the models processing the test images, we used the plugin TensorBoard from TensorFlow. As presented in Fig. 5, the results on the left show the analysis of a normal image that was submitted for classification with both RGB and Thermal models, respectively, where the identification box and its CL are presented. Additionally, in Fig. 5, the images on the right show the results after processing images with artifacts (cat and arms crossed), which shows the results of both models operating in abnormal conditions and illustrate the models' resilience to artifacts.

In this work, we used the Tensorflow lite implementation [12] by the capacity to run in edge devices, e.g. Raspberry Pi 4. In the future, we intend to explore other implementations or even improve some in order to better accurate the results of the models and other specificities of Convolutional Neural Networks, for example, works like [22] or other recent improvements using feature pyramid architecture for object detection or other new approaches for the applicability context of this work, as the architectures Resnet, R-NN-Regions with CNN features, Recurrent Convolutional Neural Network, ExtremeNet or other.

5 Conclusions and Future Work

This work is centered on the study and exploration of the complementarity of Artificial Intelligence, Machine Learning, Deep Learning, Computer Vision approaches for the identification and classification of human body parts for contactless screening systems. The main focus is to detect parts of the body in a specific position through RGB and Thermal images and not detect the segmentation or pose estimation of the human body. The proposed methodology focuses on the detection and classification of human body parts (head, torso, and arms) from both RGB and Thermal images using an implementation of the Convolutional Neural Networks through an open-source approach. The method uses a supervised learning model that can run in edge devices, e.g. Raspberry Pi 4. Using and opensource implementation and following the general methods and metrics for the model creation and validation, based on the obtained results we can conclude that the best setup for the RGB results, with the aim of the project in mind, was the third one. Because was almost infallible on the detection of the head, the torso part, compared with the other two setups, showed significantly better results and the arms part was as well above the others. We also decided to remove the arms label from the thermal dataset to preserve the general accuracy of our model. Furthermore, we also realized that the results for the thermal dataset were worse than for the RGB one, this is given the fact that the RGB was trained with superior image size. We also observed that the

placing of the camera and the distance to the actor, are fundamentals variables to obtain a good result since we had promising results with the third setup in the RGB dataset and good results in the thermal dataset. Due to the promising results achieved, we are able in the future to extend the work with the complementary approaches of human body pose estimation to improve the training strategy and network architectures on prediction accuracy using Artificial Intelligence/Machine Learning potentialities for the identification and classification of human body parts for contactless screening systems.

Acknowledgments. This work is a result of the project CoViS - Contactless Vital Signs Monitoring in Nursing Homes using a Multimodal Approach, with reference POCI-01-02B7-FEDER-070090, under the PORTUGAL 2020 Partnership Agreement, funded through the European Regional Development Fund (ERDF).

References

1. Rohmetra, H., Raghunath, N., Narang, P., et al.: AI-enabled remote monitoring of vital signs for COVID-19: methods, prospects and challenges. Computing (2021). https://doi.org/10.1007/s00607-021-00937-7
2. Vital Surveillances: The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) - China (2020). http://weekly.chinacdc.cn/en/article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51. Accessed 11 Jul 2021
3. Silva, F., Almeida, R., Pinho, P., Marques, P., Lopes, S.I.: Evaluation of a low-cost COTS bio radar for vital signs monitoring. In: 2021 IEEE International Smart Cities Conference (ISC2), Virtual Conference (2021)
4. Shavit, Y., Ferens, R.: Introduction to Camera Pose Estimation with Deep Learning. arXiv arXiv:abs/1907.05272 (2019)
5. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2938–2946 (2015). https://github.com/alexgkendall/caffe-posenet
6. Walch, F., Hazirbas, C., Leal-Taixe, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using LSTMs for structured feature correlation. In Proceedings of the IEEE International Conference on Computer Vision, pp. 627–637 (2017)
7. Fürst, M., Gupta, S., Schuster, R., Wasenmüller, O., Stricker, D.: HPERL: 3D Human Pose Estimation from RGB and LiDAR (2020). https://arxiv.org/pdf/2010.08221.pdf
8. Sárándi, I., Linder, T., Arras, K., Leibe, B.: MeTRAbs: metric-scale truncation-robust heatmaps for absolute 3D human pose estimation (2020). https://arxiv.org/abs/2007.07227
9. Véges, M., Lörincz, A.: Absolute human pose estimation with depth prediction network. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2019). https://arxiv.org/abs/1904.05947
10. Image Classification using Pytorch. https://pytorch.org/. Accessed 9 Jul 2021
11. Image Classification using Keras. https://keras.io/examples/vision/image_classification_from_scratch/. Accessed 9 Jul 2021
12. Google Tensorflow Lite webpage. https://www.tensorflow.org/lite. Accessed on 28 Jul 2021

13. Tensorflow MobileNet V2 FPNLite - Feature Pyramid Network. https://www.tensorflow.org/lite/guide/hosted_models. Accessed 13 Jul 2021
14. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In 2010 IEEE International Conference on Robotics and Automation, pp. 3108–3113. IEEE (2010)
15. Romero, J., Loper, M., Black, M.J.: FlowCap: 2D human pose from optical flow. In: Gall, J., Gehler, P., Leibe, B. (eds.) GCPR 2015. LNCS, vol. 9358, pp. 412–423. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24947-6_34
16. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. **87**(1–2), 4 (2010)
17. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
18. FLIR E54, Advanced Thermal Imaging Camera. https://www.flir.com/products/e54/. Accessed 15 Jul 2021
19. Juang, C.F., Chang, C.M.: Human body posture classification by a neural fuzzy network and home care system application. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **37**(6), 984–994 (2007)
20. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv arXiv:abs/1704.04861 (2017)
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.: MobileNetV2: inverted residuals and linear bottlenecks, pp. 4510–4520 (2018). https://doi.org/10.1109/CVPR.2018.00474
22. Ghiasi, G., Lin, T.-Y., Le, Q.: NAS-FPN: learning scalable feature pyramid architecture for object detection, pp. 7029–7038 (2019). https://doi.org/10.1109/CVPR.2019.00720
23. TensorFlow Object Detection API. https://github.com/tensorflow/models/tree/master/research/object_detection. Accessed 16 Jul 2021