— Rapport de stage de Master 2 (2020) —

Combining object detection and pose estimation to assert personal protective equipment are correctly worn

Le Bono Cédric
Lab-STICC, ENIB
c4lebono@enib.fr

Supervisor : Papadakis Panagiotis
Lab-STICC, IMT Atlantique
panagiotis.papadakis@imt-atlantique.fr

Supervisor : Buche Cédric
Lab-STICC, ENIB
buche@enib.fr

Abstract

On construction sites, workers are exposed to a wide range of hazards. To limit the risk and prevent severe work accidents, the use of PPE (Personal Protective Equipment), such as hard hat or harness, has been adopted by the industry. Monitoring the use of PPE is an inherently difficult task since its protective effectiveness depends on whether it's conformally worn. The advent of deep learning, and in particular object detection, in recent years has enabled the rise of discriminative and all-purpose PPE detectors. However, those methods most of time solely focus on detecting the equipment only. This paper presents an approach which associates object detection and human pose estimation to deduce whether the equipment is properly worn.

A video showcasing the detection of various equipment can be found at: `https://youtu.be/F7_CswI3F1g`

Contents

Introduction

According to the Occupational Safety and Health Administration (OSHA), in the USA, the construction industry suffered from 1008 worker fatalities, which represent 21.1 % of the fatalities in private industries. The two main causes of death were falls and being struck by object, respectively causing 338 and 112 deaths (OSHA, 2018). While controlling and trying to contain hazards at their origins is the best way to protect employees, this can only be done to an extent.

In this context and to alleviate the risk of accidents, regulations and safety legislation have been adopted. Notably, regarding the use of personal protective equipment (PPE). Such equipment comes in various shapes and forms to protect against a wide range of hazards. Often including items related to physical, noise or chemical protections. However, only 64% of workers wear the required equipment at all time (Farooqui et al., 2009), either by negligence or because it is inconvenient to wear.

The conventional approach to that issue is to supervise the employees and sanction them if they are not wearing their mandatory equipment. However, continuous and widespread supervision in inherently tricky and unrealistic in practice. With the advent of deep learning and computer vision, automatic and real-life equipment detectors have been developed. Using the video feed from a camera, they are able to quickly identify if the equipment is present. However, most of approaches so far focus on the detection of the equipment only. That is only partially sufficient since it is also vital to determine whether the equipment is properly worn.

This paper proposes an innovative method which combines object detection and human pose estimation. The goal is not only to detect the equipment but also assert whether it is conformally worn. At first, we will present the state-of-the-art when it comes to equipment detection. This is followed by our contribution to the problem. Our proposed solution uses both pose estimation and object detection to detect the equipment. Details of our approach are provided for the detection of hard hats, harnesses and shoes. The overall performances and results of our solution are then presented.

Related work

In the past few years, person and equipment detection on construction sites has thoroughly been investigated using

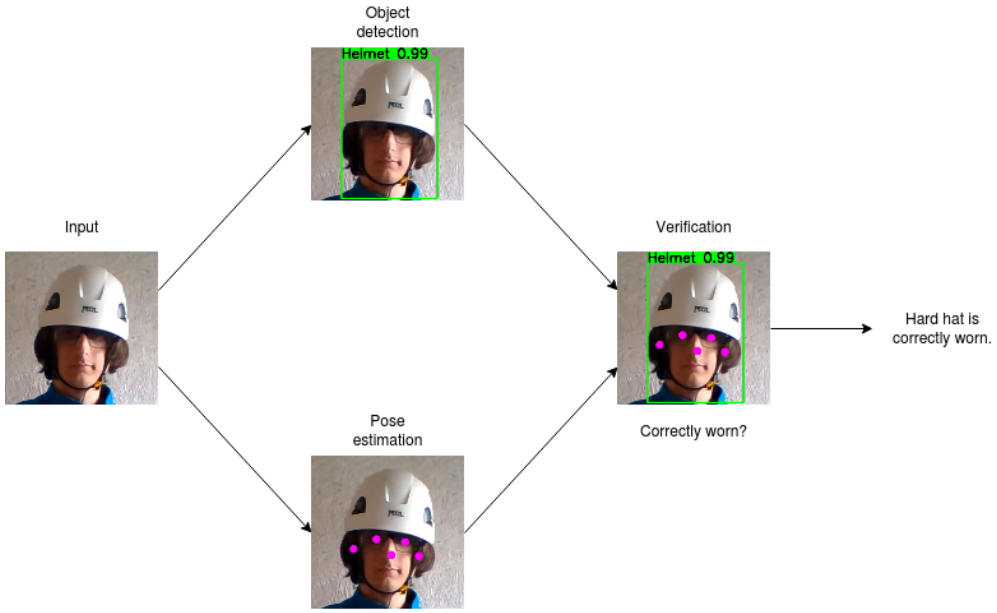

Figure 1: Architecture of our approach. Each image is sent to an object detector (YOLOV3) and a pose estimator (OpenPose). The results from both algorithms are combined to estimate whether the detected equipment detected are correctly worn.

computer vision approaches. Indicatively, worker detection (Fang et al., 2018c), harness detection (Fang et al., 2018b) and hard hats (Nath et al., 2020), (Fang et al., 2018a), (Mneymneh et al., 2017), (Wu and Zhao, 2018).

To attain their objective, those methods use state-of-the-art object detection algorithms such as Faster R-CNN (Ren et al., 2015) or YOLO (Redmon et al., 2016). Those detectors return the name of each object detected and a rectangular region containing each object called a bounding box. The main strength of those modern object detectors is that they are able to make predictions in real-time. They are thus suited for continuous monitoring of employees.

In most aforementioned approaches, the emphasis is put on the detection of the equipment. However, the detection of the equipment alone does not warrant safety. It is further required to assert whether it is properly worn and whether it is suited for the task. To our knowledge, several solutions have been proposed to check equipment is properly worn, but they remain limited. For example, performing human detection on top of the equipment detection (Fang et al., 2018b). However, this approach only checks whether a human is nearby the equipment. It only asserts whether someone is likely to wear the equipment but does not check whether the equipment is properly worn.

Regarding hard hat detection, a solution is to label the head as part of the helmet (Wu et al., 2019). A hard hat is therefore only detected if on top of someone's head. This solution works because the head has unique features which make it easy to identify. This is much more difficult when the area surrounding the equipment has more variability.

Finally, certain kinds of equipment need to be worn following strict rules as is typically the case for the *harness* equipment. A harness is considered properly worn if one of its buckle is located in the upper half of the torso. However, determining the relative position of a buckle using object detection alone is a tremendously difficult task.

Another factor to take into account is that using other or additional tools for a more robust decision-making unavoidably induces an additional computational overhead. As a consequence, this could impact the efficiency of detection and impede real-time estimation.

Methodologies

In order to detect whether an equipment is correctly worn while maintaining real-time performance, we propose a framework which combines object detection with human pose estimation.

Object detection

Ever since the introduction of Regional convolutional network networks (R-CNN) (Girshick et al., 2014), object detection has been a very active field. Faster, more accurate and easier to use detectors are nowadays made publicly available on a regular basis.

For the purpose of the proposed work, several state-of-the-arts works were studied or tested. Most notably, anchor-free single stage detector such as CornerNet (Law and Deng, 2018), CenterNet (Duan et al., 2019) or MatrixNets (Rashwan et al., 2020). They were later abandoned either because they were deemed too difficult to use or did not exhibit satisfactory performance.

On the other hand, YOLO (Redmon et al., 2016) and its later versions YOLO9000 (Redmon and Farhadi, 2017) and YOLOV3 (Redmon and Farhadi, 2018) stood out as particularly easy to use and also highly accurate. Two configurations stood out in early testing phases. Glenn Jocher's

custom YOLOV3 configuration with hyperparameter optimization greatly improved precision compared to the other regular YOLOV3 configurations (Jocher, 2018). Alexey Bochkovskiy's tiny YOLOV3 proposal which adds path aggregation networks (Liu et al., 2018). This configuration was several orders of magnitude faster then the other YOLOV3 implementations while remaining as accurate (Bochkovskiy, 2016a).

Human pose estimation

Pose estimation is a subfield of computer vision. As the name entails, its purpose is to extract the pose of humans in images or videos. For each individual in a picture, pose estimation algorithms extract the positions of multiple body joints organized into a kinematic chain.

To perform pose estimation, we used the state-of-the-art OpenPose (Cao et al., 2018) library. Compared to other pose estimation algorithms, OpenPose has the advantage of being open-source and provides real-time estimation even with multiple persons in an image. For each person in an image, OpenPose's COCO25 model return the 2D position of 25 limbs from different parts of the body.

Framework's structure

In our architecture, object detection and pose estimation are performed by respectively using the state-of-the-art YOLOV3 and OpenPose. The former is used to localize various personal protective equipment in an image, the latter to localize the body limbs. The architecture of our framework is depicted in figure 1.

Protective equipment when correctly worn are near certain body parts. By comparing the position of an equipment and the nearest most relevant body parts, it would be possible to deduce whether it is properly worn. Depending on the equipment in question, several methods could be potentially used.

In this paper, we propose two methods which can be notably used for the detection of hard hats, work shoes and harnesses. The first method is to check where the results of object detection and human pose estimation overlap. The second is to define logical conditions which need to be verified.

Overlapping

YOLOV3 returns for each detected object a bounding box. A possible approach is to check how many and which body parts are found in each box. This provides an overview of the relative position of each object with respect to one's posture.

OpenPose returns the positions of several limbs located on the head and on the feet. In this paper, this approach was used for the detection of hard hats and shoes.

An object is considered properly worn a sufficient number of relevant body limbs are located within the object's bounding box. Results of hard hats and shoes detection can be found in figure 2 and 3 respectively.

This solution is only suitable if the equipment in question can overlap with limbs which can be obtained from pose estimation.

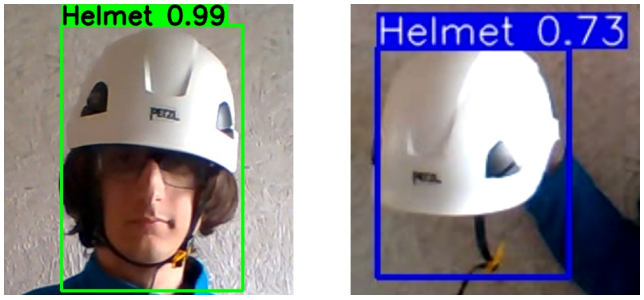

Figure 2: Hard hat correctly worn on the left (green), incorrectly worn on the right (blue).

Figure 3: Shoes correctly worn on the left (green), incorrectly worn on the right (blue).

Harness and buckle detection

Some types of equipment either do not overlap with joints or the assessment of correct use goes beyond relative position estimation. This is the case with harnesses. An harness is considered properly worn if one of its buckle is located in the upper side of the torso.

OpenPose provides 6 key points on the torso: The neck; left and right shoulders; left, mid and right hip. Checking whether the buckle is closer to neck than the middle hip would be enough. However, the detection of every body limb is not guaranteed.

The proposed approach is to assert the validity of three conditions: (a) The buckle is closer to the neck than the middle hip; (b) The buckle is closer to the left shoulder than the right hip; (c) The buckle is closer to the right shoulder than the left hip.

A buckle is considered conformally worn if the majority of these tests are positive. An example of harness and buckle detection is provided in figure 4.

Learning

Dataset

In order to train YOLOV3 to detect the equipment, a rich and well-balanced dataset of hard hats, shoes, and harnesses images is required. Unfortunately, a recurring problem with PPE detection is that no off-the-shelf dataset is available (Fang et al., 2020). A total of 1479 images of people wearing PPE were taken. Those images were taken from various conditions such as outside and inside; at close and far range; with different lightning conditions.

Additionally, 2870 images of PPE were further extracted from online videos. The purpose of those images were to obtain a wider range of PPE other than the one at our disposal.

Configuration	Metric	Helmet	Harness	Buckle	Shoes
Tiny YOLOV3 + PAN	Precision	85%	**97%**	91%	91%
	Recall	**99%**	95%	**92%**	80%
	IOU	67%	75%	69%	67%
YOLOV3	Precision	**88%**	93%	91%	**95%**
	Recall	98%	**96%**	89%	**84%**
	IOU	**73%**	**76%**	**70%**	**73%**

Table 1: Object detection: Comparison between Tiny YOLOV3 and YOLOV3.

Figure 4: Harness with two buckles, one at sternum level (green) one too low (blue)

Configuration	Metric	Result
Tiny YOLOV3	Precision	90%
	Recall	90%
	F1score	90%
	mAP@0.50	95.44%
	BFLOP	14.432
	Training (10000 iterations)	2 days
YOLOV3	Precision	92%
	Recall	90%
	F1score	91%
	mAP@0.50	96.2%
	BFLOP	139.52
	Training (10000 iterations)	5 days

Table 2: Overall results

Dataset	Source	Amount
Training	Filmed	1479
	Internet	2870
	Data augmentation	4349
	Total	**8698**
Test	Total	**779**

Table 3: Training and test datasets.

Those images were manually labeled using the Yolo-Mark tool (Bochkovskiy, 2016b).

Moreover, data augmentation was performed on the images in the training dataset. Images were flipped horizontally. This effectively doubled the size of the training dataset.

A set of 779 additional images was taken for the test dataset. The information are summarized in table. 3.

Hyperparameter optimization

As far as object detection is concerned, two configurations were kept. The tiny YOLOV3 and PAN configuration and the YOLOV3 with hyperparameter optimization. Two tables comparing the best weights obtained from both implementations can be found in tables 1 and 2.

The hyperparameters chosen for the YOLOV3 configuration can be found in table 4. Compared to Glenn Jocher's recommended parameters, the saturation, exposure, hue and jitter values were slightly increased. The higher those values are, the more the system is resilient to size and light variations.

Training was performed on computer with a GTX1060 GPU. It took approximately 2 days for the tiny YOLOV3 configuration and 5 days for YOLOV3.

Both configurations were trained for 10 000 iterations with a batch size of 64 images. Which corresponds to approximately 80 epochs. Early tests revealed that training further had little impact on the precision and recall, consequently, no further training was pursued. Graphs showing the evolution of precision and recall during training can be found in figure. 5. For both configurations, F1-score is at 89% ±2% for the last 5000 iterations.

Results

Object detection

The precision and recall obtained with both methods are fairly similar, YOLOV3 having a slight advantage over the tiny YOLOV3 and PAN implementation. It also has an higher Intersection over Union (IOU). Which means it is able to more accurately localize and evaluate the size of the objects.

Nevertheless, the differences remain fairly small and YOLO requires 10 times more floating pointer operations. As such, the trade off between speed and precision with tiny YOLOV3 is much higher.

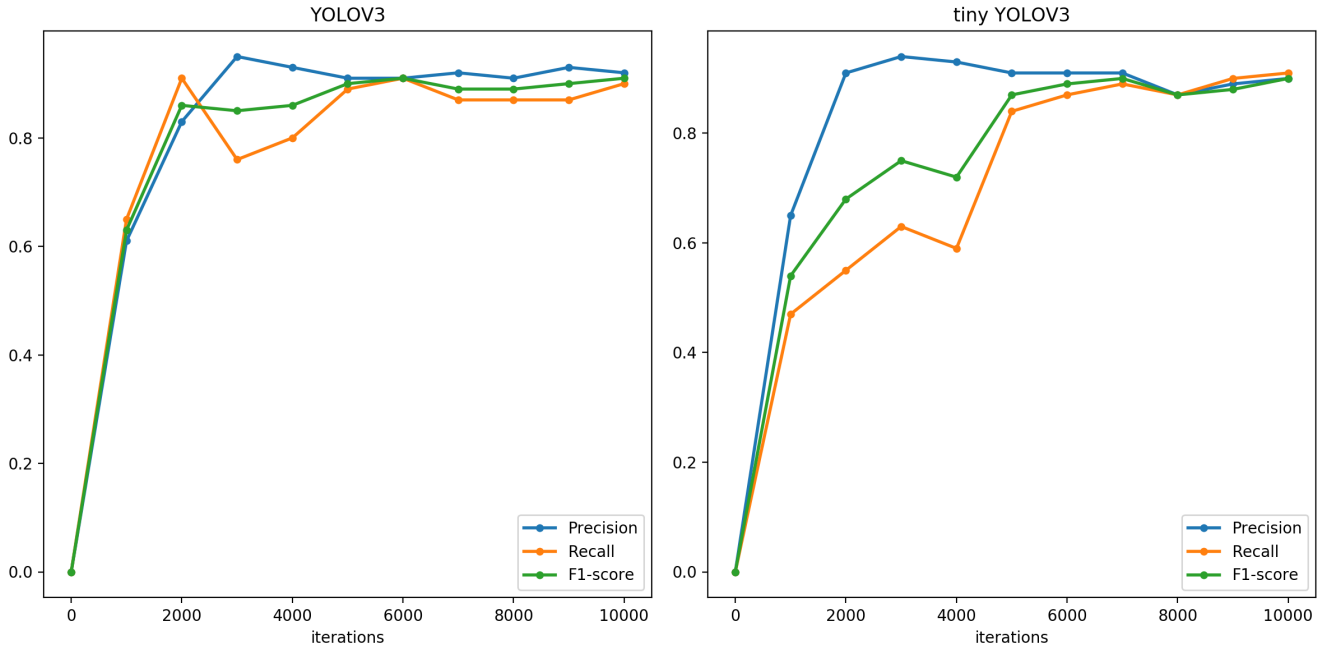

Figure 5: Evolution of precision and recall during training.

Parameter	Value
Classes	4
Width	608
Height	608
Momentum	0.949
Learning Rate	0.00261
Decay	0.0005
Saturation	2.0
Exposure	2.0
Hue	0.5
Angle	0.0
Jitter	0.5

Table 4: Selected hyperparameters for YOLOV3

Function	Time
Object detection	0.05s to 0.24s
OpenPose detection	0.5s

Table 5: Time required to process an image using a laptop with a Quadro M1000M GPU

Framework evaluation

Finally, we evaluated the performance of joint object detection and human pose estimation. For this experiment, 418 pictures of people properly wearing hard hats, harness and shoes were picked up. The purpose of this test is to demonstrate the capacity of our implementation to assess the correctness of correctly worn equipment. A qualitative result is provided in figure. 6. Both the YOLOV3 and tiny YOLOV3 configurations show high recalls.

Configuration	Helmet	Harness	Shoes
tiny YOLOV3 + PAN	93.3%	91.3%	95.6%
YOLOV3	95.8%	90.2%	94.4%

Table 6: Framework evaluation: Comparison between tiny YOLOV3 and YOLOV3.

Discussion

The combination of human pose estimation and object detection enables the extraction of complex and robust information in images. Such information would be difficult to obtain with object detection alone. However, this comes at a computing cost. Single-stage object detectors are renowned for their top notch speed. The addition of a pose estimation layer greatly reduces their efficiency.

In our experiments, object detection could be performed at about 20fps on 1280x720 videos on a laptop equipped with a Quadro M1000M GPU. Characteristically, performances dwindles to a mere 2fps with the addition of human pose estimation.

Moreover, while object detection usually focuses on precision and recall, in this work, a very important factor is also the IOU. Indeed, knowing the exact whereabouts of an equipment is mandatory to assert whether its properly equipped. A bounding box which is incorrectly positioned or with a wrong size can heavily affect the decision.

A possible improvement would be to investigate instance segmentation approaches such as Mask R-CNN (He et al., 2017). Instead of rectangular bounding boxes, this would

provide pixel-perfect localization of the objects. However, instance segmentation approaches are on the slower side of the object detectors and may not be appropriate for real-time detection once combined with pose estimation.

Extension

As far as implementation is concerned, our detector is currently running on a computer with a QuadroM1000M GPU and is available via HTTP request. By sending an image or video attached to the body of the request, our program assesses whether a person is properly equipped as depicted in figure 6.

This method has the advantage of being generic and practical. It allows computers with low processing capabilities to exploit the results of the detection.

Regarding videos, the prediction uses a voting approach. An image is extracted every 30 frames and detection is performed on it. An equipment is considered properly worn if detected as such in at least half of those images.

Conclusion and future works

In this paper, a new method is proposed for the detection of Personal Protective Equipment (PPE). Beyond the detection of equipment, our approach adds pose estimation to assert the equipment are properly worn.

Using state-of-the-art object detectors it is able to recognize PPE such as hard hats, shoes and harness with high precision and recall. Human pose estimation is used to establish whether the shoes are on ones feet, the helmets on heads and whether the harness has one of its buckle around sternum level. Such information would be difficult to extract with object detection alone.

Our current solution uses variants of YOLOV3 and obtains precision and recalls of about 90% for all objects. A possible improvement would be to experiment with the newly proposed YOLOV4 (Bochkovskiy et al., 2020). We decided not to use it as tests revealed that our GPU was not powerful enough to run both YOLOV4 an OpenPose at the same time.

Our paper focuses on those three kinds of equipment. However, it can be generalized to other equipment, pieces of garment or anything that requires the detection of objects and of people poses.

Our current implementation exposes a server to HTTP requests which performs the detection and return the objects in an image. It allows computers with little resources to exploit the result of the detection. This could for instance be used for robots which do not have access to a GPU.

A drawback of our current approach is that addition of pose estimation greatly slow downs the speed at which images can be processed. Going from 20 to 2 images per second on a laptop with a GPU.

To take it a step further, we are planning to test our proposal in simulations. By using 3D models of the equipment, we are planning to create complex environment to test the limits of our approach. This would help enhance our detector prior to putting it on the real field. A example of detection on a rendered 3D model can be found in figure. 7.

Another objective is to adjust this paper and submit it at the International Conference on Pattern Recognition 2020 (ICPR2020).

References

Bochkovskiy, A. (2016a). Tiny-yolo + pan optimization. Source code: https://github.com/AlexeyAB/darknet.

Bochkovskiy, A. (2016b). Yolo mark: Gui for marking bounded boxes of objects in images for training neural network yolo v3 and v2. Source code: https://github.com/AlexeyAB/darknet.

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. (2018). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578.

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., and An, W. (2018a). Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Automation in Construction*, 85:1–9.

Fang, W., Ding, L., Luo, H., and Love, P. E. (2018b). Falls from heights: A computer vision-based approach for safety harness detection. *Automation in Construction*, 91:53 – 61.

Fang, W., Ding, L., Zhong, B., Love, P. E., and Luo, H. (2018c). Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. *Advanced Engineering Informatics*, 37:139–149.

Fang, W., Love, P. E., Luo, H., and Ding, L. (2020). Computer vision for behaviour-based safety in construction: A review and future directions. *Advanced Engineering Informatics*, 43:100980.

Farooqui, R. U., C, P. D., Ahmed, S. M., D, P., and Azhar, S. (2009). Addressing the issue of compliance with personal protective equipment on construction worksites: A workers' perspective.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Jocher, G. (2018). Yolov3 optimization. Source code: https://github.com/ultralytics/yolov3.

Law, H. and Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750.

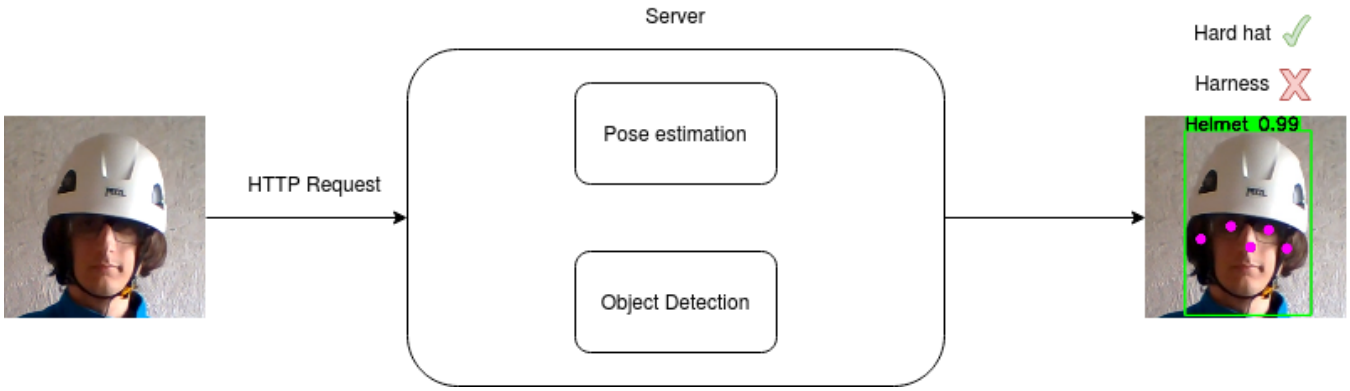

Figure 6: Server architecture

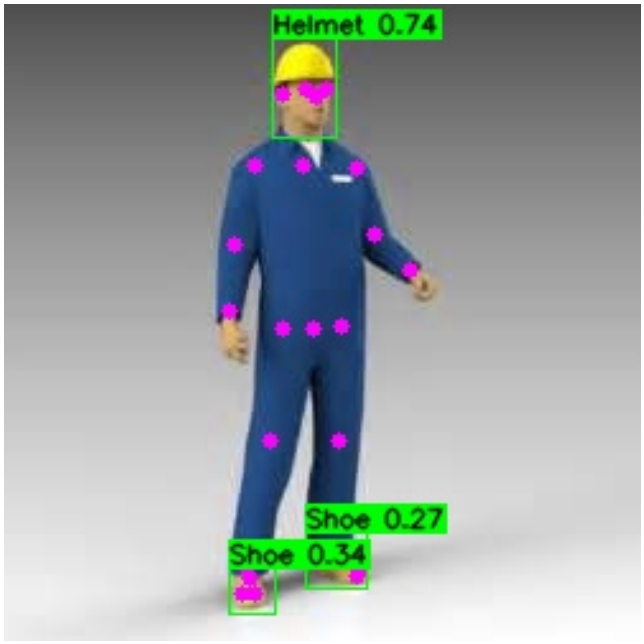

Figure 7: Detection performed on a rendered 3D model.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768.

Mneymneh, B. E., Abbas, M., and Khoury, H. (2017). Automated hardhat detection for construction safety applications. *Procedia engineering*, 196:895–902.

Nath, N. D., Behzadan, A. H., and Paal, S. G. (2020). Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, 112:103085.

OSHA (2018). Osha data and statistics. Available at: https://www.osha.gov/data/commonstats.

Rashwan, A., Agarwal, R., Kalra, A., and Poupart, P. (2020). Matrixnets: A new scale and aspect ratio aware architecture for object detection.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

Wu, H. and Zhao, J. (2018). An intelligent vision-based approach for helmet identification for work safety. *Computers in Industry*, 100:267–277.

Wu, J., Cai, N., Chen, W., Wang, H., and Wang, G. (2019). Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. *Automation in Construction*, 106:102894.