Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/38/20230561

Two-stage helmet security detection model combining CNN and MTCNN algorithm

Jiangchuan He

Faculty of Engineering, The University of Hong Kong, Hong Kong, 999077, China

u3577194@connect.hku.hk

Abstract. Physical safety is of utmost importance for workers in an industrial and construction environment. Fatalities related to the violation of wearing a safety helmet account for a large proportion of occupational deaths. Hence, the requirement of a high-efficiency automatic system to monitor and promote wearing helmets is of vital importance, reducing hours from manual monitoring, facilitating efficient education and enforcement campaigns that increase industrial safety. In this paper, a combined Convolutional Neural Network (CNN) algorithm and Multi-Task CNN (MTCNN) to monitor violations of safety helmets is proposed. All images are sourced from industrial and construction locations, undergoing various image processing techniques before model detection. The proposed model includes the MTCNN locating individuals' heads from the picture data, followed by a CNN assessment to determine helmet wearing. Under the same evaluation criterion, the result, where an obvious detection accuracy improvement is shown, illustrates the effectiveness of the proposed model in the comparative study. Therefore, the research demonstrates the effectiveness of incorporating multi-scale feature extraction techniques for enhancing model performance. This establishes the groundwork for future research within this domain, facilitating the creation of more advanced helmet detection or other relevant items recognition systems that can be utilized in real-world applications.

Keywords: CNN, MTCNN, Helmet Detection, Image Processing.

1. Introduction

Nowadays, video surveillance and recognition technology are extensively utilized, including criminal tracking and auto ticket checking by facial analysis, traffic violations monitoring, and in healthcare sector like conservation of the elderly. Nevertheless, a significant number of industrial accidents occur due to the violations of safety helmets. During the last decade, there were over 65,000 cases per year that involved head injuries resulting in missed workdays in the workplace. In 2022, 1,020 individuals suffered fatal workplace injuries due to severe head injuries [1]. As the leading type of head injuries, Traumatic brain injuries (TBI) accounts for around 30% of all injury deaths in industrial areas [2]. Despite the safety helmet standard afforded to workers, TBI incidents, particularly in construction, still occur, with 2,200 fatalities recorded during the last decade. Nevertheless, a survey on occupational injuries launched by the Bureau of Labor Statistics (BLS), showing that 84% of all workers lack of head protection who suffered TBI. Monitoring safety helmet violations manually in industrial environments is consuming and ineffective [3].

DOI: 10.54254/2755-2721/38/20230561

Consequently, it is crucial to implement an automatic surveillance system to supervise and urge people to wear helmets in industrial environment. Machine learning and its applications in image processing have been significant due to its computational ability and effective identification and detection on target items. However, implementing the model for real-time detection presents challenges in many scenarios, particularly when performing applications in complicated environments or on low-performance devices. The state-of-the-art deep learning one stage object detection methods performs satisfactorily with limited device performance like Single Shot Multi Box Detector (SSD) [4], which is a developed Convolutional Neural Network (CNN) based model and Yolo [5] (You Only Look Once) are useful. To run smoothly on low-performance devices, both models sacrifice accuracy and potential to handle complex problems. Moreover, SSD models have a relatively high performance only on large objects compared to small objects [6]. As a consequence, SSD are frequently used in conjunction with other lightweight feature extractors, which utilize different usage of depth wise separable convolution [7], contributing to various performance advantages and limits. Based on characteristics above, a method, detecting safety helmets effectively by optimizing the performance of basic CNN model on smaller size objects, is naturally came up. The structure of two connected CNN models is proposed. The first CNN model is used to recognize vague head or person positions from images, and the next one focuses on the helmet recognition problem.

The objective of this study is to develop a reliable helmet detection model using deep learning techniques such as CNN, Multi Task Convolutional Neural Network (MTCNN). Particularly, CNN, MTCNN, were developed for different tasks. MTCNN is used to locate person's head in a complex picture, then, cropping the head images of persons and passing it to CNN algorithm. CNN is used to do some simple helmet detection on the relatively simple pictures only focusing on person's head, and it generates output, judging whether the person is wearing a helmet. The simple data shows faces and helmets with a minimum of influence from the background. By combining these two steps, the model can detect person in the input picture or video and then determining whether a person is wearing helmet or not. In conclusion, the aim of developing helmet detection models using deep learning techniques like CNN, MTCNN is to automate and simplify the identification of helmet wearing behavior in images. This project addresses the need for efficient and accurate helmet detection tools and a long-term fight against the accidents caused by not wearing helmets in construction area by protecting workers' safety and reducing labour costs. This would significantly enhance safety measures in the industrial sector by greatly reducing incidents of traumatic brain injuries resulting from not wearing safety helmets.

2. Methodology

2.1. Dataset description and preprocessing

The dataset used in this study is sourced from Kaggle containing pictures about workers [8]. Construction area and other industrial environment, where wearing safety helmet is mandatory, are major source of pictures. There are 5000 images in total (Figure 1), with bounding box annotations for these 3 classes: Helmet, Person, Head. For training and analysis, images containing people with and without helmets are categorized, since helmet detection is the primary goal. With different angles of persons and also different sizes of the heads and helmets, the training set maintains a rich variety, which automatically enhances its complexity, conductive to efficient model training [9]. The pictures have slight pixel differences and each of them are around 2 MB size.

Figure 1. Sample pictures from the dataset.

DOI: 10.54254/2755-2721/38/20230561

2.2. *Proposed approach*

The main objective of this study is to develop a reliable helmet detection model using deep learning techniques such as CNN and MTCNN. This method is built upon the solid basis of CNN, which is well-known, combined with the MTCNN developing from CNN and enhancing its multi features handling abilities, making it performs quite outstanding in locating persons head in a complicated background. Detection carried out by MTCNN positioning persons in pictures is arranged before the judgment from CNN on whether the person is wearing the safety helmet. There are three major steps in the model: a) MTCNN for head detection b) Resizing the images of individuals' heads and passing them to the CNN c) Implementing CNN algorithm for helmet detection. The reason for using this two-model structure is to improve the overall performance of the detection model, especially handling the cases with complicated background, which could drastically affect the accuracy of the system. Combining these technologies, the model will be able to accurately detect if individuals are wearing helmets in a given scenario, reminding people to wearing helmet and further reducing the risk of injuries. Figure 2 below illustrates the structure of the system

Figure 2. Structure of the model.

2.2.1. CNN. The CNN, a type of artificial neural network, excels at image recognition and is thus widely utilized. Its structure is grid-like, just like an image, and preserves some key factors of the input images in relevant tasks. Consequently, it performs exceptional on picture detection. CNN replace general matrix multiplication with a mathematical operation known as convolution in one or more of their layers. There are three types typical CNN layers: Convolution, Pooling, and Fully Connected [10]. Furthermore, a different activation function is applied to the layer to improve its performance (Figure 3).

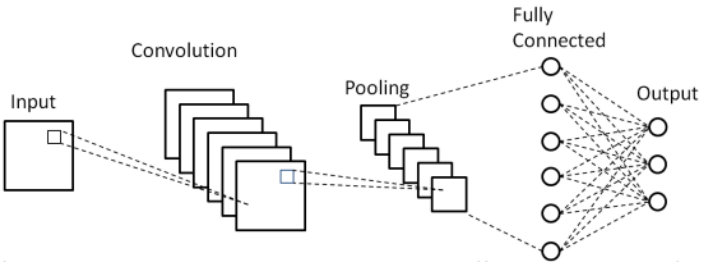

Figure 3. An example of CNN structure.

The Convolution Layer extract primary visual picture features by simulates simple cells, utilizing local connection and weight sharing techniques. Local connection denotes the link between convolution layer and preceding layer, connecting each neuron on these adjacent layers. Weight sharing entails locally connecting neurons to the prior layer with a shared value evaluating connection strength. This technique reduces the network's training parameters. Activation function is applied in computing parameters in layers, aiming to add some no-linear factors, helping the model to solve some complicated problems since linear models lack the necessary expressiveness. The Pooling Layer compresses the input feature map, reduces its size, and simplifies the network's computational complexity, extracting the primary features. Fully connected layers are responsible for serving as the "classifier" in the entire CNN. It establishing the association between the represented feature vectors and the labels that correspond to them.

DOI: 10.54254/2755-2721/38/20230561

2.2.2. MTCNN. MTCNN is used as a head locating algorithm in the model. The network structure of the algorithm consists of three layers, P-net, R-net and O-net. It is developed from CNN, and performs better on detecting multiple objects in complex pictures. Therefore, it is used to locate person's head, then, cropping the head images of persons and passing it to CNN algorithm for further detection. There are four essential steps for face recognition: one preprocessing step for scaling and three processing steps in MTCNN for face region detection (Figure 4).

Figure 4. Brief Flowchart for MTCNN.

The input image is first scaled by means of the image pyramid in preprocessing step. After scaling, faces can be extracted regardless of size. Pnet network is followed and performs a coarse screening of the image retrieved from the image pyramid to obtain a number of tentative bounding boxes for husky faces. To obtain a fairly accurate bounding box for the face, it then performs a non-maximum suppression operation. These rough preliminary bounding boxes are passed to the Rnet refinement network for the next step. Note that the non-maximum suppression (NMS) operation is implemented to obtain the maximum value in the local region, suppress other lower values, and output the best detection result. To confirm that it is indeed a human face, Rnet identifies objects within the inner region of these coarse face boxes. At the same time, Rnet adjusts the height and width of the boxes and continues to apply NMS to produce improved boxes that go beyond the previous step. Rnet then passes these relatively good boxes on to the Onet. Onet evaluates the interior of these frames to identify human faces. It then improves the height and width of these boxes to better correspond to the size of the human face and ultimately executes NMS, culminating in the final prediction of human face boxes (Figure 5).

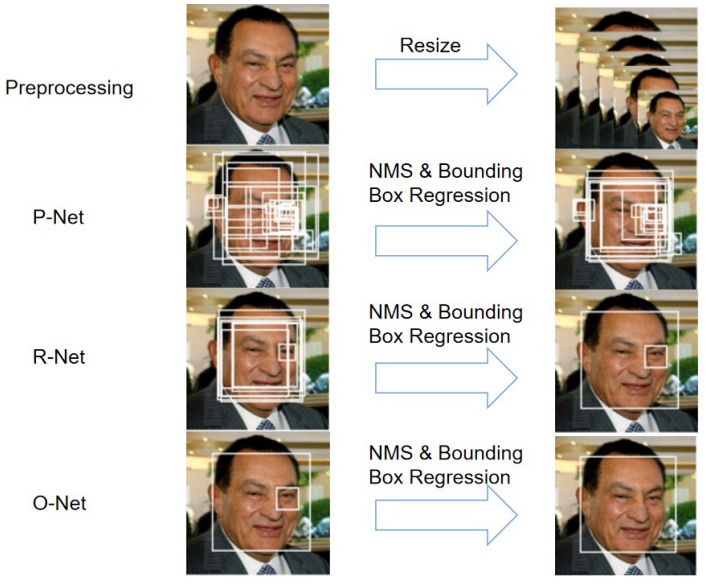

Figure 5. MTCNN working steps with sample picture interpretation.

2.3. Implemented details

The task of constructing the model is divided into four stages, which contain preprocessing, building, testing and improving, and applying to better organize and implement the functions (Figure 6).

DOI: 10.54254/2755-2721/38/20230561

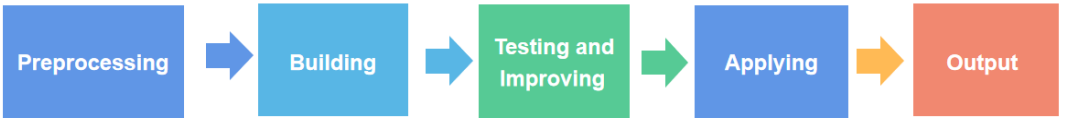

Figure 6. Implementation steps.

In the preprocessing stage, input RGB pictures are standardized and resized to 128*128 pixels. These pictures are randomly sorted to prevent the model from getting stuck in local optima or errors, which can be caused by focusing on incorrect features. A train-test split is performed, with 4000 pictures in the training set and 1000 in the testing set out of 5000 pictures. The model is built utilizing previously stated knowledge, utilizing various layers such as convolutional, max pooling, and fully connected layers. Preprocessed data from the previous step is used to train the model. During testing and improving stage, results are obtained from the model using the testing set. The accuracy rate, loss rate, time cost, and confusion matrix illustrating the number of incorrectly identified with helmet or without helmet pictures are all considered. After training the helmet detector, helmet detection model is applied to data set. Some images are sent to the model, which determines whether the person in the image is wearing a helmet, and outputs some pictures with boxes pointing helmets.

3. Result and discussion

In the conducted study, a two-stage model containing the MTCNN and CNN was utilized to perform helmet detection from a collection of 5000 images, each labeled with a specific bounding box annotation.

As can be seen from the Figure 7, the MTCNN+CNN model achieves 99% and 96% accuracy on train set and test set respectively, also sharing a similar trend. The accuracy rate experiences a steep incline until the eighth epoch for both sets, followed by a gradual increase until the tenth epoch for the test set. During tenth to twentieth (approx.) epochs, accuracy increased slowly on both sets, but then stabilized with very insignificant changes, remaining at a high level.

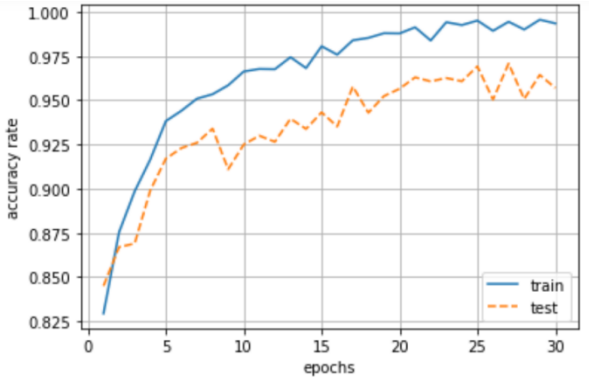

Figure 7. The result of proposed model, MTCNN+CNN model.

A comparative analysis of the accuracy of the two-stage model against that of the standalone CNN model is provided in Figure 8. Both figures have number of epochs and accuracy rate on the X and Y axis respectively, evaluating the result of training. The Figure 9 shows the result of MTCNN+CNN model at the early stage before optimization. As depicted in Figure 8, the accuracy of the standalone CNN model exhibits wavelike rise before stabilizing at a lower final accuracy rate, around 81%. The orange line represents the same data, accuracy rate of the MTCNN+CNN model on test set, in both Figure 7 and 8. Additionally, higher initial accuracy of the combined model demonstrating that it provides a more efficient solution for cold-start scenarios. The superior performance of the MTCNN+CNN model is attributed to the multi objects detecting ability facilitated by the MTCNN structure. Merging the MTCNN with CNN allows for the combination of detecting primary key

DOI: 10.54254/2755-2721/38/20230561

information, persons' heads this case, and further specialized determination on weather wearing the helmet.

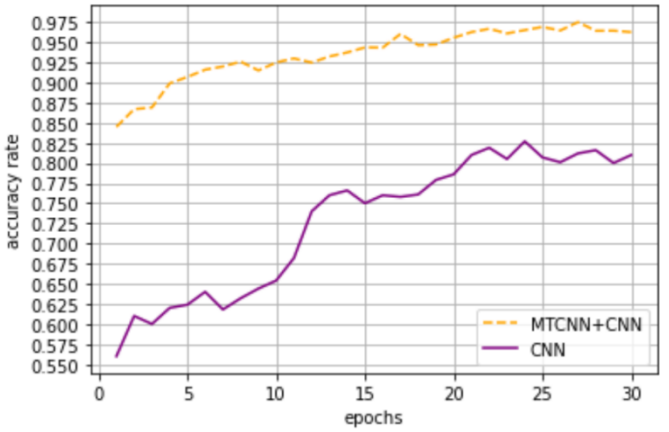

Figure 8. Comparison between the MTCNN+CNN with CNN model on the same test set.

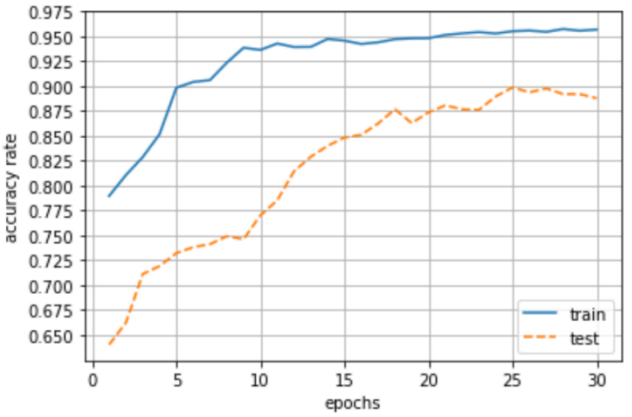

Figure 9. The result of early stage MTCNN+CNN model before optimization.

Therefore, both sub-models, MTCNN and CNN, can be made to focus on a part of the big problem, that is, the identification of wearing a helmet in a complex environment is broken down into two subproblems: human head locating and the determining helmet is wore or not. Preprocessing of the MTCNN backbone is a basic step in improving the overall performance of the model. The solution of two sub-models handling two subproblems separately reduces the training complexity of a single model and improves the overall training accuracy, results in a 15% improvement. Key steps in improving the overall performance of the model through tuning learning rate, optimizing the structure of Rnet, Pnet, Onet and the application of enhancements such as bag of tricks [11], Random Erasing is implemented, giving an 8% improvement from around 90% (Figure 9). In summary, the combination of MTCNN and CNN to construct a two-stage model resulted in a significant improvement in helmet detection performance. The precision line chart intuitively illustrates this improvement, while also highlighting areas that require further improvement. Future research could consider exploring alternative combined network architectures or advanced training strategies to further optimize emotion recognition accuracy. Some recognition results are shown below, green boxes denote persons with helmets, while red boxes referring not wearing helmets (Figure 10).

DOI: 10.54254/2755-2721/38/20230561

Figure 10. Left side of figure shows persons with helmets, right side shows persons without helmets.

4. Conclusion

This paper proposes the two-stage helmet security detection model to identify the violation of wearing safety helmets in industrial environment, reducing manual monitoring costs. The model first using MTCNN to get predicted persons' head positions. Adjacently, the images are further cropped, with head position annotated, and passed to helmet judgement step carried out by the CNN algorithm. The proposed two stage MTCNN plus CNN model is compared with CNN for helmet detection accuracy, on the identical settings and training data, with performances further evaluated on the same test data. The proposed model effectively recognizes persons and determines helmet wearing in images, achieving an accuracy of 96%. This outcome signifies a remarkable 15% enhancement in comparison to using the standalone CNN model, which only reaches 81% accuracy on the same test data set. The experimental results have illustrated that the approach improves the basic CNN model efficiently on helmet detection task. For future research, tuning different hyper-parameters by training multiple models to testing different hyper-parameters and different layer numbers and sizes to optimize the model. More advanced deep learning architectures will also be examined to enrich the model's interpretive ability.

References

[1] Kamboj A Nilesh P 2020 Safety helmet detection in industrial environment using deep learning Proceedings of the 9th International Conference on Information Technology Convergence and Services (ITCSE 2020)

[2] Maas A I R et al 2022 Traumatic brain injury: progress and challenges in prevention clinical care, and research The Lancet Neurology

[3] Van Z Adriaan H Van D 1985 Residual complaints of patients two years after severe head injury Journal of Neurology Neurosurgery & Psychiatry 48(1): pp 21-28

[4] Liu W et al 2016 SSD: single shot multibox detector Computer Vision ECCV 2016 - 14th European Conference, Amsterdam 2016 Proceedings pp 21–37

[5] Redmon J Divvala S K Girshick R B and Farhadi A 2016 You only look once: Unified real-time object detection 2016 IEEE Conference Computer Vision and Pattern Recognition (CVPR) pp 779–788

[6] Huang J Fathi A Rathod V Fischer I and Sun C 2017 Speed/accuracy trade-offs for modern convolutional object detectors Computer Vision and Pattern Recognition arXiv:1611.10012

[7] Howard A Wang W Zhu M Weyand T Chen B and Kalenichenko B April 2017 MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications Computer Vision and Pattern Recognition arXiv:1704.04861

[8] Dataset https://www.kaggle.com/datasets/andrewmvd/hard-hat-detection

[9] Shorten C and Khoshgoftaar T M 2019 A survey on Image Data Augmentation for Deep Learning Journal of Big data 6: p 60

[10] Phung Rhee 2019 A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets Applied Sciences 9: p 4500

DOI: 10.54254/2755-2721/38/20230561

[11] Zhang Z et al 2019 Bag of freebies for training object detection neural networks arXiv preprint arXiv:1902.04103