Assembly training system on HoloLens using embedded algorithm

Yujin Qin*, Shuxia Wang, Qiang Zhang, Yao Cheng, Jiaxu Huang, Weiping He
Cyber-Physical Interaction Lab, School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, Shaanxi Province, China
* Corresponding author: qinyujinnpu@foxmail.com

ABSTRACT

In this article, we demonstrate an implementation on Microsoft HoloLens, deep learning supported in the context of object detection. The main aim of the training system is to create the more accurate object detection model for Augmented Reality using deep learning models for image recognition directly on the HoloLens 2. In terms of the object detection approach, a deep learning model called YOLOv5 has been used for the implementation of this system. This article uses the Windows ML API to implement machine learning in augmented reality applications. A simple and easy method of drawing lines between specified 2D coordinates on a canvas is proposed. The module division and development steps of the development of augmented reality training system are given. Our system provides the annotation of augmented object detected and its bounding box via HoloLens. It allows to detect the new object in a few milliseconds. Preliminary results show a great rate of object detection and reasonable detection time.

Keywords: Augmented Reality; Object detection; deep learning; Microsoft HoloLens; detection time

1. INTRODUCTION

In the era of intelligent manufacturing, the digital, The high-performance assembly of major and complex products such as aero-engines has attracted extensive research by domestic and foreign scholars. Due to the large number of parts and the complicated information of paper process documents, the stationary rocket engine is mainly assembled manually at present. In order to ensure the quality and efficiency of engine assembly, assembly training and cognitive operation of new employees are very important. However, the cognitive training of aero-engine assembly mainly adopts the traditional mode of " experienced people teach newbies ", which requires large labor costs and learning cycles.

With the updating of computer software and hardware, mixed reality technology has begun to be widely used in education, industry, medical and other industries. Mixed reality technology which can deeply integrate information of different dimensions breaks through traditional space limitations and interaction methods. Mixed reality includes virtual reality and augmented reality. The simulation training system developed by virtual reality technology [1] has the characteristics, such as immersive learning features, high security and so on. However, most of the existing virtual training systems can only achieve independent interactive learning, which cannot meet the needs of operators for scene perception. Augmented reality solves the above problems by ingeniously integrating virtual objects with real scenes. In addition, object detection has good results in object recognition of assembly parts and perception of assembly environment. The integrated training system which uses technologies related to augmented reality and object detection develops a new way of human-computer interaction。It can provide theoretical and technical support for problems existing in the learning training and visual guidance.

This article designs and develops a training system for solid rocket motor mandrel based on augmented reality and deep learning. The system solves the problem of long training time and low efficiency in on-site assembly and training of complex products such as aero-engines. Moreover, we achieve a high degree of fusion between machine learning models and augmented reality devices, which significantly improves field perception in assembly cognitive training.

2. RELATED WORKS

At present, the target detection based on deep learning has surpassed the traditional target detection algorithm, and has greatly improved the accuracy and recognition efficiency. Deep learning algorithms for object detection include CNN, RCNN, fast-RCNN, and YOLO, etc. The second generation YOLO2 of the YOLO algorithm in 2017: Better, Faster, Stronger [2], which realizes the recognition of various types of objects. The original neural network structure is improved

Third International Symposium on Computer Engineering and Intelligent Communications (ISCEIC 2022),
edited by Xianye Ben, Proc. of SPIE Vol. 12462, 124620H · © The Authors. Published under a
Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2660940

by replacing the fully connected layer with the convolutional layer. In 2018, Redmon proposed YOLOv3 [3], which uses the FPN structure for multi-scale feature extraction. In order to improve the detection effect of mAP and objects, the algorithm replaces Softmax with Logistic to form a deeper network level and multi-scale detection. In 2021, Li Yangfan et al. [4] proposed an improved YOLOv4 spatial infrared weak target detection method, with an average recognition accuracy rate of over 93.25% and a detection speed of 38.99ms/frame. In 2021, Song Xin of Dalian University of Technology [5] used the X-YOLOv3 network structure to identify parts of gears and bearings in the workshop in a simulated workshop. The results show that this method can reduce the work pressure of operators.

AR represents an innovative tool that can ensure the assembly efficient and correct transfer of knowledge. HoloLens device is recognized as the best solution for augmented reality applications, and image recognition and processing has become an inevitable trend in the augmented reality environment. In 2016, Zhang Lijuan [6] of Huazhong University of Science and Technology combined mobile phones and computers, and introduced augmented reality technology into the equipment maintenance system of the Expo Center. This method greatly facilitates the management of the equipment by the staff. In 2017, Zhang Shupeng[7] applied AR virtual interaction technology to the work of knowledge popularization, which effectively improved the learning efficiency of users. In 2018, Feng Danlin[8] of the Civil Aviation University of China used HoloLens to develop a mixed reality interactive demonstration system for civil aviation engines, which can intuitively understand the structure and principles of the engine. However, this interactive system is only a virtual engine model and cannot be used in combination with real engines and virtual engines. In 2019, Derakhshani et al. [9] proposed a technique of adding local activation, which effectively improves the MAP of the YOLO detector without affecting the computational speed. In 2020, Zhou Xiang of Fuzhou University and others [10] proposed an augmented reality system based on the image recognition BRISK-SURF algorithm. The system has high recognition accuracy for transformations such as illumination and occlusion, which further improves the computing speed of the system. In 2021, Zeng Xiao of Sichuan University [11] proposed a target detection system based on TCP/IP remote communication. The system breaks through the traditional interaction method, but there are problems of high information delay and data loss.

In order to further reduce the information delay in the interaction process and improve the on-site perception in the assembly process, this paper uses the YOLO algorithm to develop a solid rocket motor mandrel on-site training system. The system implements the independent target detection function of HoloLens2 through deep learning algorithm processing on the Microsoft HoloLens device. The training system can update the enhanced information corresponding to the assembly parts according to the detection results, so that the user can interact more intuitively.

3. SYSTEM DESIGN

This section will describe the main working process of the system, the overall scheme of the solid rocket motor mandrel field training system is shown in the figure1. In terms of hardware, this article uses the processor of Microsoft HoloLens2 (Qualcomm Snapdragon 850) to perform the calculation of the machine model; Meanwhile, the HoloLens2 camera acts as an input module for object detection in the real world. The camera mounted on the front of the device allows the app to intuitively detect what the user is seeing. Developers can obtain device metadata by accessing and manipulating the camera. In the training system, we need access to the HoloLens research mode streams, as well as OpenCV to enable real-time processing of sensor frames streamed from the Microsoft HoloLens device in a C# Unity environment using IL2CPP Windows Runtime Support.

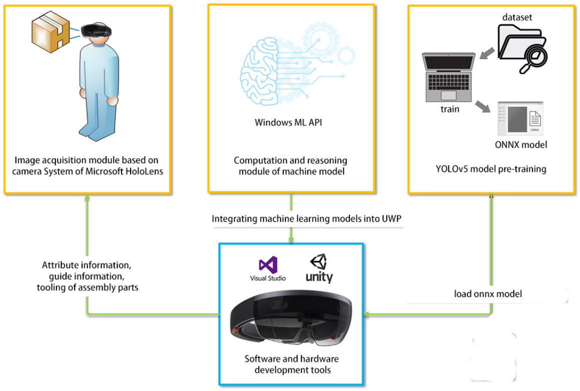
Image acquisition module based on camera System of Microsoft HoloLens
Windows ML API
Computation and reasoning module of machine model
dataset
train
ONNX model
YOLOv5 model pre-training
Integrating machine learning models into UWP
Attribute information, guide information, tooling of assembly parts
Visual Studio
unity
load onnx model
Software and hardware development tools

Figure 1. Software and hardware system diagram.

The training system construction process mainly includes two stages: machine learning model pre-training and system detection and recognition.

In terms of machine learning model pre-training, we use the automatic acquisition program we created to collect and frame the on-site assembly parts firstly. Secondly, performing Mosaic data augmentation on raw images and creating a dataset. This operation can enrich the background of the detected object and improve the detection effect of small target parts. Finally, we use the YOLOv5 network to perform transfer learning and training on the features of the assembled parts.

The operation process of the training system is shown in Figure 2. Firstly, capturing the HoloLens2 media frame and using the algorithm to obtain the latest video frame. Then we pass the video frame into the main program and wait for the asynchronous command. Secondly, the main program loads the pre-trained machine learning model and label files. The latest video frame is evaluated as the network input, and the network prediction results are output. Finally, according to the prediction results, the UI interface guidance information of the HoloLens holographic image is updated.

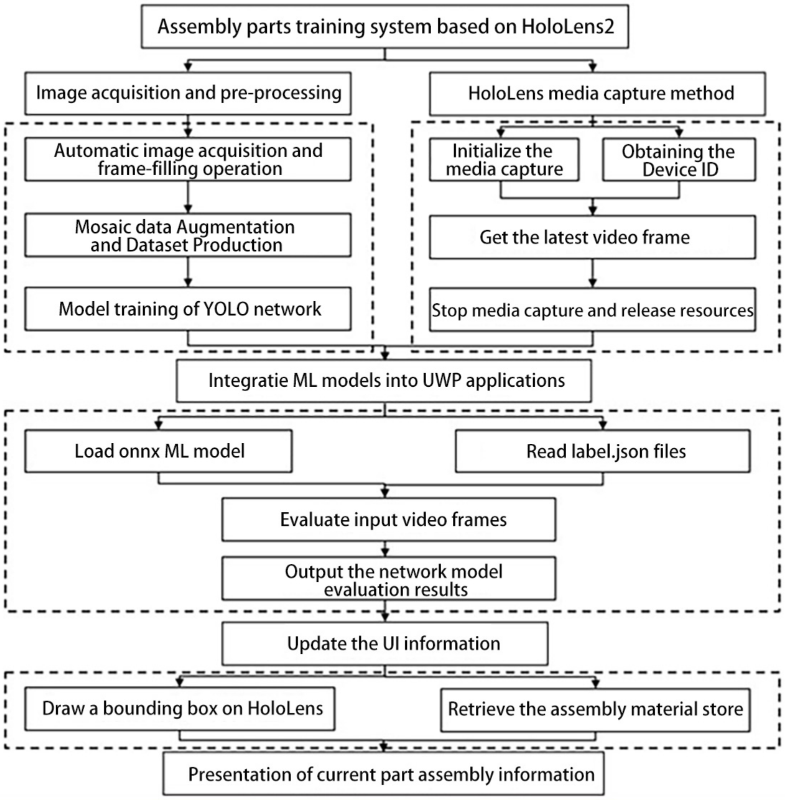
Assembly parts training system based on HoloLens2
Image acquisition and pre-processing
HoloLens media capture method
Automatic image acquisition and frame-filling operation
Mosaic data Augmentation and Dataset Production
Model training of YOLO network
Initialize the media capture
Obtaining the Device ID
Get the latest video frame
Stop media capture and release resources
Integratie ML models into UWP applications
Load onnx ML model
Read label.json files
Evaluate input video frames
Output the network model evaluation results
Update the UI information
Draw a bounding box on HoloLens
Retrieve the assembly material store
Presentation of current part assembly information

Figure 2. System Structure Diagram.

4. SYSTEM COMPONENTS

In order to facilitate the subsequent upgrade of the training system, according to the idea of modular design, the on-site training system of the solid rocket motor core mold in this paper can be divided into four parts: 1: Model training based on YOLOv5n. 2: Integrating machine learning models into UWP applications. 3: Capture of HoloLens media frames. 4: the display and update of GUI interface .

4.1 Model training based on YOLOv5n

The data set is made by using the pictures collected on site, and the model is trained on the parts data set. In order to run better on HoloLens devices, this article uses the pre-trained weights of the lightweight model YOLOv5n, with a size of 3.9MB. Secondary training of YoloV5N-CBAM model is carried out on solid rocket motor parts dataset. The relevant parameters in model training are set as follows:

Table 1. Relevant parameters of model secondary training.

Item	Parameter	Item	Parameter
Img_size	640*640	epoch	200
Batch_size	4	momentum	0.937
Class	12	1r0	0.01
Warmuo_epochs	3.0	Warmup_bias_1r	0.1

In the table, the ing_size represents the input image size; epoch represents the number of model iterations; batch_size represents the number of images for each training of the model. Momentum means bias correction for the optimizer, class refers to the category of engine parts, 1r0 represents the initial learning rate, warmup_epochs represents the warmup training period, and warmup_bias_lr represents the learning rate for the warmup training period. In the process of model training, the performance of the computer graphics card is directly related to the number of images trained in each iteration and the training time.

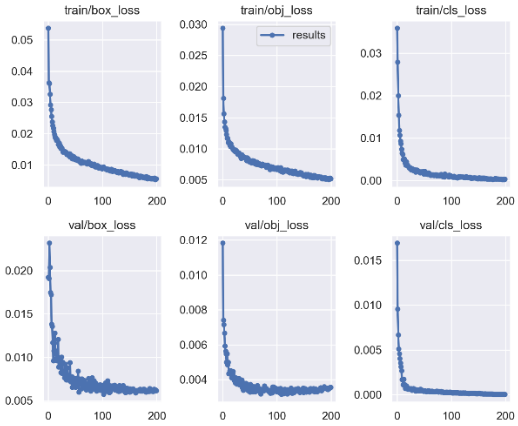

Figure 3. Loss function change diagram.

It can be seen from the figure that the frame regression loss value in the training set has the fastest decline rate in the warm-up learning cycle. As the number of iterations increases, the decreasing speed of the frame regression loss value gradually slows down until it becomes stable. The initial value of the target detection loss value is very small, and it slowly decreases with the increase of epoch until it falls below 0.01 and then slowly converges. The classification loss value only needs 20 epoch to start to converge. In the validation set, the three types of loss values decreased rapidly and gradually became stable in the first 20 epochs. The model training ends after 200 epochs, and we get 3.7MB of weight data for both.

4.2 Integrating machine learning models into UWP applications

This module converts a pre-trained machine learning model to an Open Neural Network Exchange (ONNX) model via the tf2onnx tool and adds it for integration into the application. Finally, an executable application is generated and deployed to the HoloLens device, as shown in Figure 4.

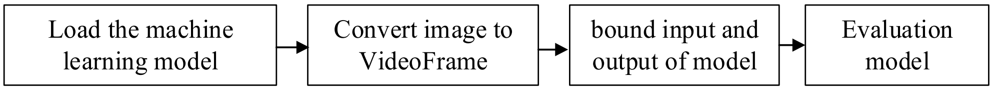
Load the machine learning model
Convert image to VideoFrame
bound input and output of model
Evaluation model

Figure 4. Integrated development process.

Windows ML is a high-performance and reliable API, which is used for hardware-accelerated ML inference on Windows devices. Therefore, this article uses the Windows Machine Learning API to deploy TensorFlow models in Windows applications. The process includes: (1) We can access the ONNX model through the LoadModel method and store it in memory; (2)Convert the incoming original image to VideoFrame format; (3)We need create a session for binding the model, and bind the model to the specified device, so that the model can be executed on the specific device of the computer.A machine learning model has input and output features that are used to pass information into and out of the model. (4) Evaluate the model and communicate its predictions.

4.3 Capturing of HoloLens media frames

This module uses MediaFrameReader in conjunction with MediaCapture to obtain media frames from one of the available sources, including color, depth, infrared cameras, etc. This function is designed to provide raw input to programs that process media frames in real time.

Use the MediaCapture class to capture photos and videos, and use MediaFrameReader to process media frames. Since the captured image needs to be used immediately in the app, the image should be captured into a SoftwareBitmap. Finally, creating a method keeps getting the latest video frame from the media frame reader.

4.4 The display and update of GUI interface

After obtaining the prediction results of the deep learning model for the assembly parts, this article uses HoloLens to interactively display the prediction results. The display of the prediction results of the parts can further improve the operation effect of the training system. In order to achieve this purpose, it is first necessary to use HoloLens to determine the boundary area of the recognized object. Subsequently, data is indexed in the unique assembly parts material library so that identification information can be updated in real time.

Since there is no appropriate method in Unity to draw a rectangle (or bounding box) on a canvas, this section extends the Texture2D methods in Unity. We use the Unity Vector2.Lerp method, which is used to linearly interpolate between two vectors by a specified increment. Therefore, this method achieve the aim of drawing lines between specified 2D coordinates on a canvas and construct a bounding box, complete with label and confidence. Among them, the incremental frac of linear interpolation can be expressed as:

$$frac = \frac{1}{\sqrt{(x2 - x1)^2 + (y1 - y2)^2}} \tag{1}$$

During the generation of the bounding box, HoloLens utilizes the embedded processor to index a library of materials specific to the assembly components. The material library includes attribute information for assembly components, assembly guide information, and associated parts. After obtaining the specific object information, the label and UI information is updated, thereby improving the interaction efficiency during the training process and expanding the dimension of the information.

5. EXPERIMENTS AND RESULTS

5.1 Recall and Precision

In order to evaluate the recognition effect and operating efficiency of the training system, this article evaluates the recognition effect of assembly parts through the recall rate and precision rate of the training model. The recall rate and precision rate can be expressed as:

$$R[i] = \frac{\sum_{j=1}^{i} TP[j]}{\sum_{j=1}^{i} TP[j] + \sum_{i=1}^{i} FN[j]}, j = 1, 2, ..., n \tag{2}$$

$$P[i] = \frac{\sum_{j=1}^{i} TP[j]}{\sum_{j=1}^{i} TP[j] + \sum_{j=1}^{i} FP[j]}, j = 1, 2, ..., n \tag{3}$$

First, sort the n test results in descending order of confidence. Second, TP means that the correct part recognition frame is predicted. Among them, the FN array represents the positive samples that are predicted to be negative, and the FP array represents the negative samples which are predicted to be positive.

Since there are many parts involved in the training system in this paper, the AP value of each part is calculated by Formula (4) to judge the recognition efficiency of each part. The AP value of a certain category can be expressed as:

$$AP_i = \frac{1}{n} \sum_{r\in\{0,0.1,\ldots,1\}}^{i} \max_{\tilde{r}:\tilde{r}\geq r} P(\tilde{r}), i=1,2,\ldots,n \tag{4}$$

The map which is mean of AP for all categories can be expressed as:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{5}$$

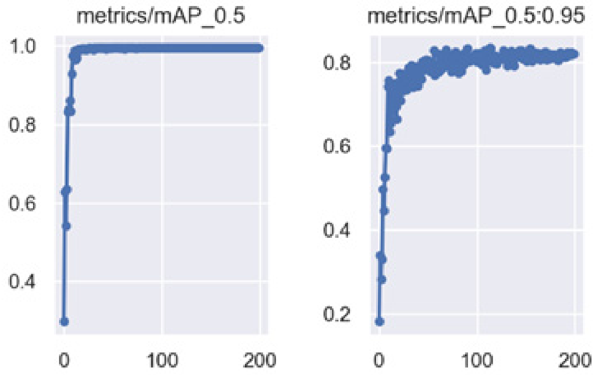

Figure 5. Change trend graph of mAP_0.5 and mAP_0.5:0.95 during training.

The training results in this article are shown in Figure5. As the number of iterations increases, the values of mAP_0.5 and mAP_0.5:0.95 increase rapidly in the first 20 iteration step cycles, and eventually they tend to be stable. As can be seen, the maximum value of mAP_0.5 is finally stable at 0.98. At the same time, the maximum value of mAP_0.5:0.95 is stable at around 0.82. The results of each category of Recall and Precision obtained after training are shown in Table 2.

Table 2. Recall and precision for each object.

Classification	**Recall**	**Precision**
Exhaust	78.2%	96.8%
Mandrel I	71.9%	93.5%
Mandrel II	62.5%	95.6%
Upperpanel	61.4%	96.8%
Bolt	75.9%	98.2%
Shell	86.3%	97.3%
Headring	69.4%	92.1%
Downpanel	85.1%	93.1%
Frock	76.8%	98.6%

Further, the detection time of the system directly reflects the sensitivity and overall performance of the training system. This article deploys and runs the system on the HoloLens2 device. To use the Unity timer function to systematically time the inference process of the experimental network model. Under the current enhanced equipment conditions, the final detection time of the integrated system is shown in Table 3.

Table 3. Detection time for each object.

Classification	**Detection time (/s)**
Exhaust	0.362
Mandrel I	0.387
Mandrel II	0.420
Upperpanel	0.298
Bolt	0.421
Shell	0.364
Headring	0.358
Downpanel	0.402
Frock	0.319

In order to test the robustness of the system, this paper conducts ensemble experiments on 1000 kinds of machine learning models. The experimental results s how that the average detection time is about 0.475s. This time indicates that the method performs well for a certain number of categories.

6. CONCLUSION AND FUTURE WORK

In this paper, we built an augmented reality system that supports object detection and visual human-computer interaction by combining the latest augmented reality device HoloLens2 and the advanced target detection algorithm YOLOv5. This method embeds the machine learning model into the HoloLens device through the Windows ML API. The experimental results show that the method proposed in this paper has high recall rate and high precision rate on target detection, and the accuracy rate on the test set and the average accuracy rate of all categories have better performance. At the same time, using the HoloLens2 augmented reality helmet gets rid of the traditional human-computer interaction methods and provides theoretical and technical support for visual display. However, this paper lacks control experiments in the design of experiments, and the changes to the algorithm itself are small. The experiments will continue to be improved in future work.

REFERENCES

[1] Jiang, L. Development of Virtual Reality Training System. 2021. Virtual, Online, Singapore: Association for Computing Machinery.

[2] Redmon, Joseph, and Ali Farhadi. ”YOLO9000: better, faster, stronger.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271. 2017.

[3] Redmon, Joseph, and Ali Farhadi. ”Yolov3: An incremental improvement.” arXiv preprint arXiv:1804.02767. 2018.

[4] Yangfan Liu, Lihua Cao, Ning Li,Yunfeng Zhang. Space infrared weak target detection based on YOLOv4 [J]. LCD and Display,2021,36(04):615-623.

[5] Xin Song. Research on Parts Recognition in Simulated Workshop Based on Convolutional Neural Network [D]. Dalian University of Technology,2021.

[6] Lijuan Chen, Hanbin Luo, Hongyan Xin.Application of Augmented Reality Technology in Equipment Maintenance System of Wuhan International Expo Center [J]. Construction Technology,2016,45(06):37-40.

[7] Shupeng Zhang , Wenjunn Hou, Ximeng Wang . Research on the design of popular science knowledge learning method based on augmented reality virtual-real interaction [J]. Packaging Engineering,2017,38(20):48-55.

[8] Danlin Feng. Development of mixed reality interactive demonstration system for civil aviation engine based on HoloLens [J]. Digital Technology & Application,2018,36(11):152-154+156.

[9] Derakhshani, Mohammad Mahdi, Saeed Masoudnia, Amir Hossein Shaker, Omid Mersa, Mohammad Amin Sadeghi, Mohammad Rastegari, and Babak N.Araabi. “Assisted Ex-citation of Activations: A Learning

Technique to Improve Object Detectors." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9201-9210.2019.

[10] Xiang Zhou, Lilin Tang, Ding Lin. Natural feature virtual-real registration method based on binary robust invariant scale keypoint-accelerated robust feature[J].Journal of Computer Applications,2020,40(05):1403-1408.

[11] Xiao Zeng. Research on target detection technology based on HoloLens2 [J]. Computer software and computer applications,2021(14):92-95.