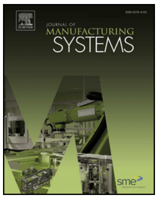
Contents lists available at ScienceDirect

Journal of Manufacturing Systems

journal homepage: www.elsevier.com/locate/jmansys



Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing

Ze-Hao Lai[a], Wenjin Tao[a,*], Ming C. Leu[a], Zhaozheng Yin[b]

[a] Department of Mechanical and Aerospace Engineering, Missouri University of Science and Technology, Rolla 65409, USA
[b] Department of Computer Science, Stony Brook University, Stony Brook, NY 11794, USA

ARTICLE INFO

Keywords:
Intellegent manufacturing
Augmented reality
Deep learning
Object detection
Region-based Convolutional Neural Networks (R-CNN)
Mechanical assembly

ABSTRACT

Quality and efficiency are crucial indicators of any manufacturing company. Many companies are suffering from a shortage of experienced workers across the production line to perform complex assembly tasks. To reduce time and error in an assembly task, a worker-centered system consisting of multi-modal Augmented Reality (AR) instructions with the support of a deep learning network for tool detection is introduced. The integrated AR is designed to provide on-site instructions including various visual renderings with a fine-tuned Region-based Convolutional Neural Network, which is trained on a synthetic tool dataset. The dataset is generated using CAD models of tools and displayed onto a 2D scene without using real tool images. By experimenting the system to a mechanical assembly of a CNC carving machine, the result of a designed experiment shows that the system helps reduce the time and errors of the given assembly tasks by 33.2 % and 32.4 %, respectively. With the integrated system, an efficient, customizable smart AR instruction system capable of sensing, characterizing requirements, and enhancing worker's performance has been built and demonstrated.

1. Introduction

In the Industry 4.0 era, consumer needs towards products of high quality, high complexity, and mass customization have been growing at an increasingly fast-moving pace. Many companies are seeking solutions that could increase the efficacy. However, the state of having a shortage of experienced workforce has been a critical problem while employers are facing the rapid transition. According to a Honeywell report [1], 78 % of the modern technology is considered important, 65 % of the technological advances are restrained by the outdated work styles, and 38 % of the current workers are actively looking for different positions. This has reflected the urgent needs for the system update in workforce training. Also, the quality of products plays a vital role as the difficulty of assembly increases, e.g., a jet engine is comprised of more than 10,000 individual parts assembled together. As reported by GE [3], the company has lost millions of dollars each year because nuts and hoses that seal fluid lines are not fastened properly for the jet engine, which leads to an unnecessary cost from repairing, not to mention the safety of passengers.

Therefore, to improve the productivity, the ability to sense, monitor, characterize, and support workers for highly complex assembly has become even more imperative, especially when conducting unpleasant, unsafe, exhausting tasks. In order to remain competitive, engineers and researchers have been attempting for solutions toward intelligent manufacturing by applying emerging technologies such as Virtual Reality (VR) and Augmented Reality (AR) technologies, Artificial Intelligence (AI), Internet-of-Things (IoT) and industrial digital twins [4,39].

1.1. Related work

In this paper, we focus on applying AR and AI technologies on workforce training for mechanical assembly tasks. The related work is categorized into industrial applications and academic research of AR in workforce training, and AI technologies.

In industry, many leading manufacturing companies have noticed the potential and started piloting AR technologies, which has been successfully utilized in various fields ranging from the medical area [5] to the assembly line. The significant reductions of errors, time, and training requirements have been measured and proven by the Augmented Reality for Enterprise Alliance (AREA) of Boeing, which stated "This has tremendous potential to minimize errors, cut down on costs and improve product quality" [3]. GE witnessed the improvement in productivity and efficiency by implementing AR. Honeywell also proved the success in worker training with the usage of AR [1]. Although AR rendering for assembly has been demonstrated for its

* Corresponding author at: 311 Engineering Research Lab, Missouri University of Science and Technology, 500 W. 16th Street, Rolla, MO 65409, USA.
E-mail address: w.tao@mst.edu (W. Tao).

https://doi.org/10.1016/j.jmsy.2020.02.010
Received 3 July 2019; Received in revised form 23 February 2020; Accepted 26 February 2020
Available online 04 March 2020

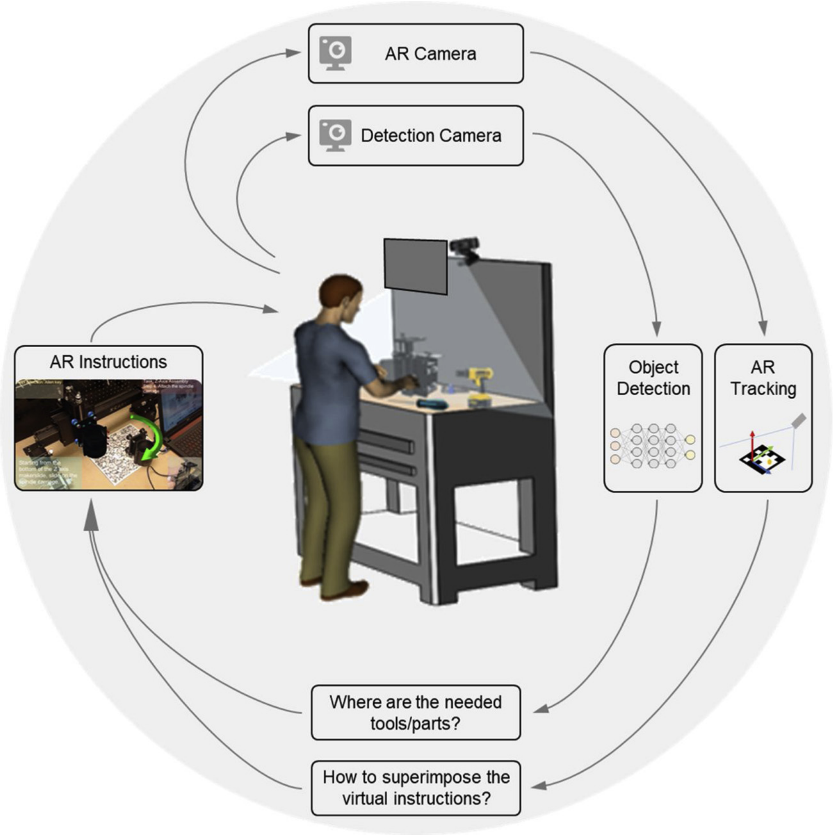

AR Camera
Detection Camera
AR Instructions
Object Detection
AR Tracking
Where are the needed tools/parts?
How to superimpose the virtual instructions?

Fig. 1. Overview of the proposed smart AR instructional system.

promising potential, industries still look for real-world solutions aiming at increasing productivity by minimizing the assembly time and errors with the assistance of AR. Commercialized AR supported devices are usually too heavy to wear and too expensive. This may induce concerns in industry use, especially at the assembly line.

In the perspective of academic study, many researchers have emphasized the benefits that AR could bring to the industry. Tao et al. [6] discussed the state-of-the-art VR/AR technologies for assembly simulations including modeling, sensing, and interaction. Esmaeilian et al. [49] concluded an important asset that data visualization helps decision making and reduces time latency in production. Caudell et al. [7] proposed an AR application for manual manufacturing processes. Azuma et al. [8] presented the potential of AR with a head-mounted display (HMD). Zhang and Kwok [40] developed an AR interface for additive manufacturing. Over the past decade, more and more research regarding engineering assembly has become popular as engineers have applied AR to different engineering scenarios [9–13]. For AR training, Webel et al. [14] raised the problem of handheld device during assembly. Recently, Tao et al. developed a training & assistant system for worker-centered intelligent manufacturing with self-awareness and active-guidance [39]. Leu et al. [15] pointed out the research efforts needed to improve the realism of virtual assembly, such as high-fidelity dynamic graphic displays, low-cost sensor fusion techniques, haptic devices, and multi-modal rendering capabilities [16]. Also, the lack of natural interactive mechanisms between the assembly operators, the assembly of components, and the instructions rendered needs to be addressed. Werrlich et al. [17,18] presented an overview of evaluations using AR training, by identifying the current limitations pertaining to high similarities of existing designed experiments that need to be improved.

With the advancement of technologies in AI, especially in machine learning and deep learning, machines are now able to perform as good as, or even better than humans in various tasks, such as image recognition [41], image segmentation [42], hand gesture classification [19,20], speech recognition [43], language translation [44], robotics [45], and intelligent vehicle [46–48]. For human activity recognition, Davide et al. [21,22] successfully recognized basic motions using signals captured from a smartphone by extracting features using classification. Ward et al. [23] proposed a strategy for recognizing assembly motions. Tao et al. [24] developed a Convolutional Neural Networks (CNN) model to recognize the worker activity using IMU and sEMG signals captured from an armband. Al-Amin et al. [25] used a Kinect sensor to perceive the worker's activity for workforce modeling and management. Overall, the number of research papers for deep learning [26] methods, such as pattern recognition using CNN [27], R-CNN [28,29] for object detection, have been growing rapidly. Additionaly, more research aimed at learning features from synthetic datasets using data augmentation has been validated [30,31] for 3D object and pedestrian detection, which identifies the utilities when training data is limited. For industry use, Wang et al. [50] presented an extensive survey and feasibility by comparing traditional machine learning to deep learning in manufacturing application, reflecting a growing trend in implementing integrated intelligent manufacturing systems to leverage human operation [51]. On the other hand, deep learning approaches can be as well applied to tool wear assessment modeling [52], indicating another possibility of monitoring equipment status for manufacturers.

1.2. Overview of the proposed system

In this study, we propose a smart instruction system with the support of AR and deep learning-based tool detection, which is intended to improve the worker performance through assistive instructions. To develop the system, two cameras are applied to capture the frame data of a working area from different perspectives. The reason of why we use two cameras is that, considering real-life shop floor assembly in which production line would separate working area and tools to prevent parts from being mixed with tools, two cameras for capturing two perspectives are setup as required. The augmented view is rendered via an onsite display and the frame data for tool detection are sent to a fine-tuned deep learning network for predictions. While the predicted results are transmitted to the AR system, it superimposes the assembly instructions accordingly. Fig. 1 illustrates the overall system workflow. The main contributions of this paper are summarized as follows:

1 An Augmented Reality instructional system integrated with Faster R-CNN for mechanical assembly is proposed.
2 A synthetic tool dataset is developed by using data augmentation with CAD models and was successfully deployed for detecting real tools.
3 The system reduces the assembly time and errors significantly in a manual assembly task compared with conventional method of using the paper manual instruction.

The remainder of the paper is organized as follows. Our proposed system is detailed in Section 2, in terms of multi-modal augmented reality, tool detection, and system integration. The experimental setup and evaluation metrics are described in Section 3. Experimental results and discussion are presented in Section 4. Finally, Section 5 provides the conclusions of this study.

2. Materials and methods

The smart AR instructional system contains two major components, a multi-modal augmented reality interface and a tool detection module, which are interfaced and integrated through an internet protocol.

2.1. Multi-modal augmented reality

To provide visual instructions, an AR interface which consists of various information in different modalities, including texts, graphics, and animations, has been developed. With multi-modal AR instructions, workers can directly sense the physical environment when following the AR instructions. Fig. 2 illustrates the proposed multi-modal AR interface.

To realize the multi-modal AR display, a marker is attached to the workbench so the webcam can capture the information of the patterns from the marker for feature recognition and tracking, so as to superimpose the computer-generated (CG) data for a composite view. To achieve the data overlaying process, an effective camera pose estimation approach for coordinate transformation is adopted, which is discussed below.

First, features (corners) are extracted for target recognition. Corners are regions with significant variations in intensity, which can be detected using a sliding window to measure the variation of intensity. The equation of the sliding window is:

$$E(u, v) = \sum_{x,y} w(x, y)[I(x + u, y + v) - I(x, y)]^2 \tag{1}$$

where $w(x, y)$ is the window function (e.g., uniform box window or Gaussian weighted window) on the position x and y of an image. $I(x, y)$ is the intensity at (x, y). u and v are represented as the shifting distances in x and y directions, respectively. To find the features that yield the highest $E(u, v)$, the second term needs to be maximized, meaning the largest change in intensity $E(u, v)$. Fig. 3(a) presents a texture-rich image selected for the marker in this study due to its large number of features (corners). Fig. 3(b) illustrates the corners on the marker which are detected as features for target recognition and tracking, which can be utilized for developing a 3D world coordinate system in camera pose estimation.

After extracting features, the camera pose estimation using homography transformation is applied. Fig. 4 illustrates the pipeline of realizing an AR effect. To augment a computer-generated data onto a scene, the homography transformation estimates the camera pose with a projection matrix. In this method, the calculation is initiated based on the pinhole model assumption of the RGB webcam. The projection matrix is an integrated matrix that combines an intrinsic matrix of the camera and an extrinsic matrix comprised of a 3 × 3 rotation matrix and a 3 × 1 translation vector.

The equation of the camera pose estimation using homography is as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathrm{H} \begin{bmatrix} X_W \\ Y_W \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ 1 \end{bmatrix} \tag{2}$$

where (u, v) are the coordinates in the 2D image plane of the display. (X_W, Y_W) uses the world coordinate system based on the detected features of the target marker. The homography transformation between the two coordinate systems, as illustrated in Fig. 5, is

$$\mathrm{H}= \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{3}$$

where h_{11}...h_{33} are its parameters. With the aid of the derived projection

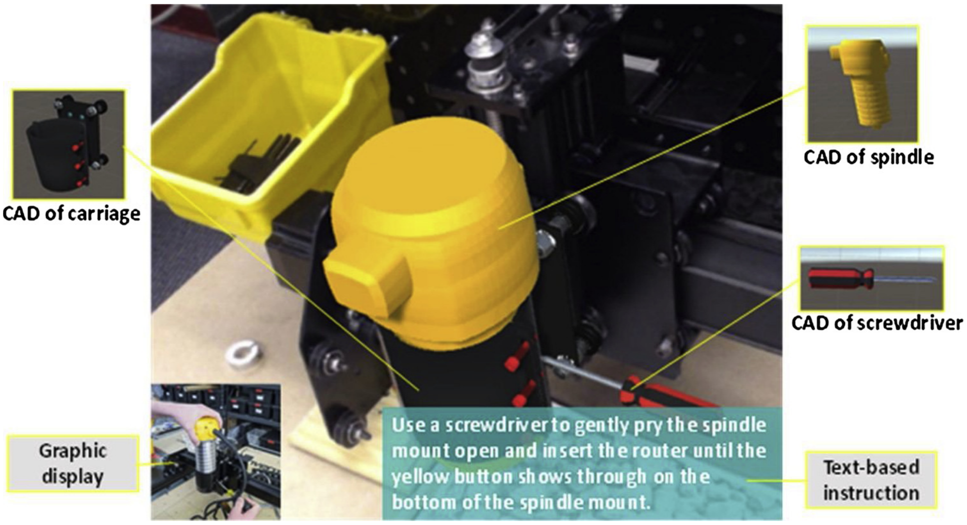

CAD of carriage
CAD of spindle
CAD of screwdriver
Graphic display
Use a screwdriver to gently pry the spindle mount open and insert the router until the yellow button shows through on the bottom of the spindle mount.
Text-based instruction

Fig. 2. The proposed multi-modal AR instruction system. Multiple types of instructions are rendered through displays including text, graphics, and 3D animations.

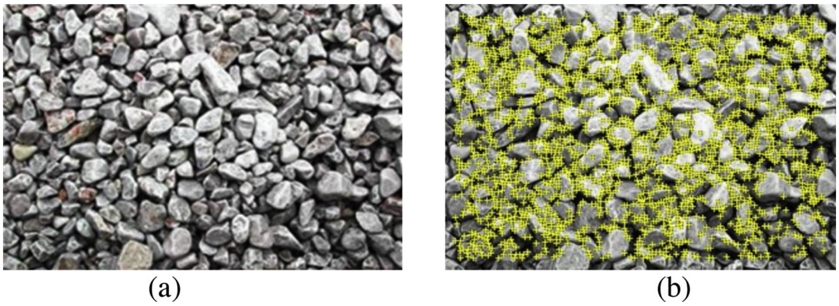


Fig. 3. A texture-rich marker and the detected feature points. Corners represent the regions which have the highest change in intensity in all directions, which are highlighted with '+'.

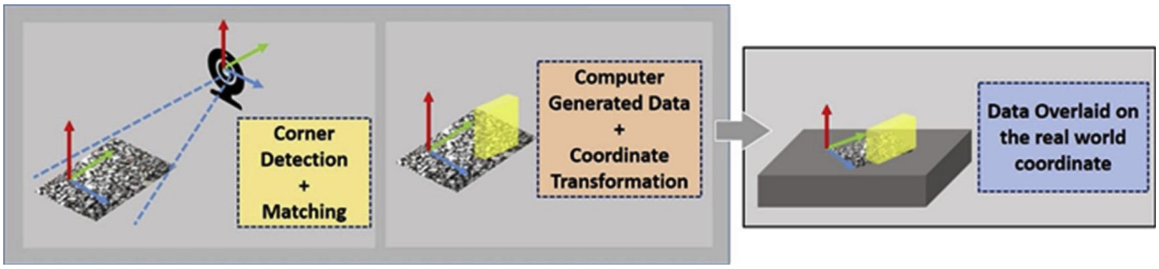

Corner Detection + Matching
Computer Generated Data + Coordinate Transformation
Data Overlaid on the real world coordinate

Fig. 4. The pipeline of realizing AR. An attached marker and the features within the pattern are detected and recognized. The local coordinate system based on the markers is generated for estimating the relation between the marker and the camera. The computer-generated data can be overlaid once the estimation is complete.

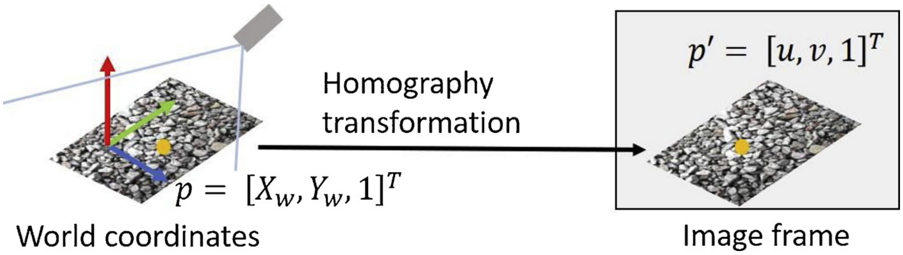


Fig. 5. The homography transformation between the two coordinate systems.

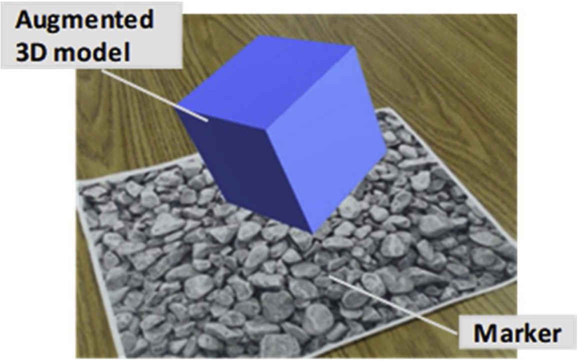


Fig. 6. An example frame of an AR effect. A 3D cube is overlaid on the target marker using the homography transformation.

matrix, a 3D model can be overlaid onto a 2D image plane with transformed coordinates as a composite view. As an example, an AR composite view using the homography is shown in Fig. 6.

2.2. Tool detection

During a manual assembly operation, to efficiently assemble every component is essential. To prevent from misusing tools, a deep learning-based tool detector is proposed to help workers locate correct tools with visual guidance. The detection is realized with a webcam mounted on the top of the workbench to capture the video frames of the working area. The workflow of the proposed tool detector is illustrated in Fig. 7.

2.2.1. Detection with faster R-CNN

This section describes the tool detector developed using a Faster Region-based Convolutional Network (R-CNN) [32]. The detection incorporates a webcam that captures 2D frames of the working area and feeds them into the model, which is trained on an annotated (labeled) dataset for target classification and localization. Given a video frame or an image, the detector outputs both classification and localization results of tools. Faster R-CNN is developed based on CNN, which has been validated as a robust network for different levels of feature extraction. Fig. 9 summarizes the overall workflow of the Faster R-CNN architecture.

2.2.1.1. Region proposal network. After extracting features using a CNN, an assigned Region Proposal Network (RPN) [32] is responsible for producing high-quality proposals (bounding boxes) with multiple scales and aspect ratios and sliding them across the convolutional feature map to detect objects. Fig. 8 illustrates the generated region proposals with multiple scales and aspect ratios.

Instead of feeding multiple unselected proposals computed from the external approach such as the selective search method [34], RPN computes the possibility of whether if there is an object in the proposal. It will pass forward to object classification and bounding-box regression

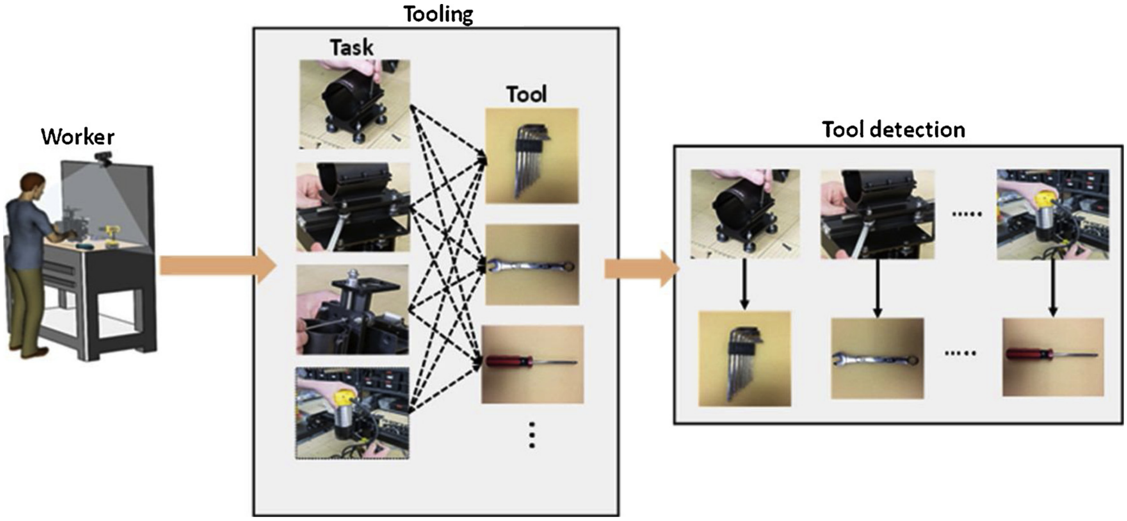


Fig. 7. The system workflow of the tool detector in the assembly.

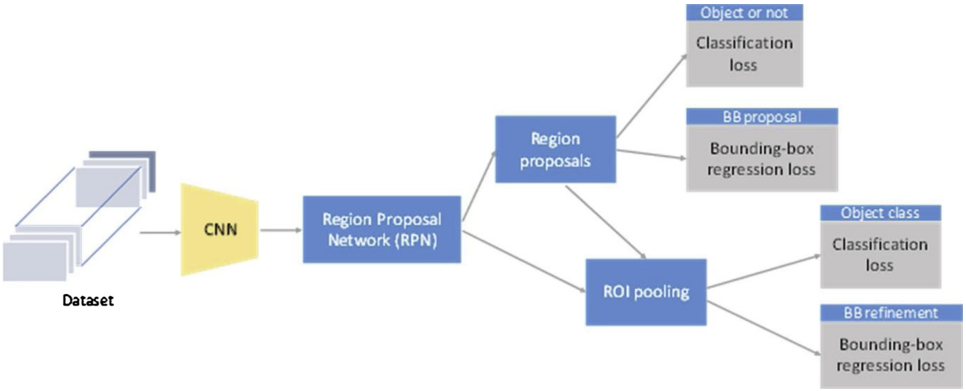


Fig. 8. Region proposal network (RPN). An RPN generates multiple proposals and slides through the convolutional feature map output from a CNN.

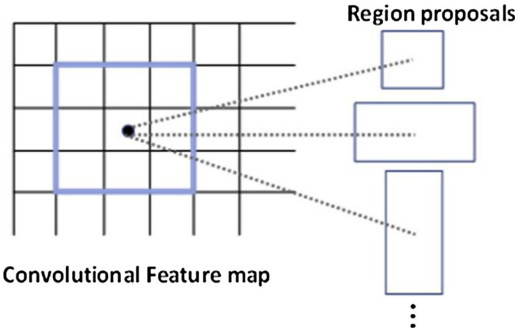


Fig. 9. Illustration of the Faster R-CNN architecture.

only if there is an object detected inside the proposal.

2.2.1.2. Object classification. After computing the object score using RPN, an ROI (Region of Interest) pooling layer is inserted to reduce the computation of the network by down-sampling the spatial size of the parameters. The classification of the detected object in the bounding-box is achieved using a Softmax function to predict classification scores over 5 classes of tools as follows:

$$P(y_i x_i) = \frac{\exp(S_i)}{\sum_{k=1}^{5} \exp(S_k)} \tag{5}$$

where $P(y_i x_i)$ is the predicted probability of a given image x_i and S_i, $i \in [1,5]$, is a 5-dimensional score vector representing the five different classes of tools. These five probability scores are normalized between zero and one as confidence scores that sum to one.

2.2.1.3. Bounding-box regression. For the bounding-box regression of the detected object, the bounding-box regressor is adopted from [28]. During training, N pairs of ground-truth boxes G and proposed boxes P are defined as training inputs, which are denoted as $\{(P^i, G^i)\}_{i=1,\dots,N}$, where $P^i = (P_x^i, P_y^i, P_w^i, P_h^i)$ represents the pixel coordinates of the center, width, and height of P^i. The ground-truth boxes are also represented as $G^i = (G_x^i, G_y^i, G_w^i, G_h^i)$. The training process is to learn the transformation and map the proposed box P to G, which is denoted as four functions: $d_x(P)$, $d_y(P)$, $d_w(P)$, $d_h(P)$. After the transformation is learned, the predicted box $\widehat{G}$ can be generated by using the following transformations:

$$\widehat{G}_x = P_w d_x(P) + P_x \tag{6}$$

$$\widehat{G}_y = P_h d_y(P) + P_y \tag{7}$$

$$\widehat{G}_w = P_w \exp(d_w(P)) \tag{8}$$

$$\widehat{G}_x = P_h \exp(d_h(P)) \tag{9}$$

In the above equation, each $d_*(P)$ (where * is one of x, y, h, w) is denoted as $w_*^T \phi(P)$, where $\phi(P)$ is modeled as a linear function of the features of a proposal and w_* is a vector of learnable model parameters [28].

2.2.1.4. Loss function. As the training process begins, each image frame

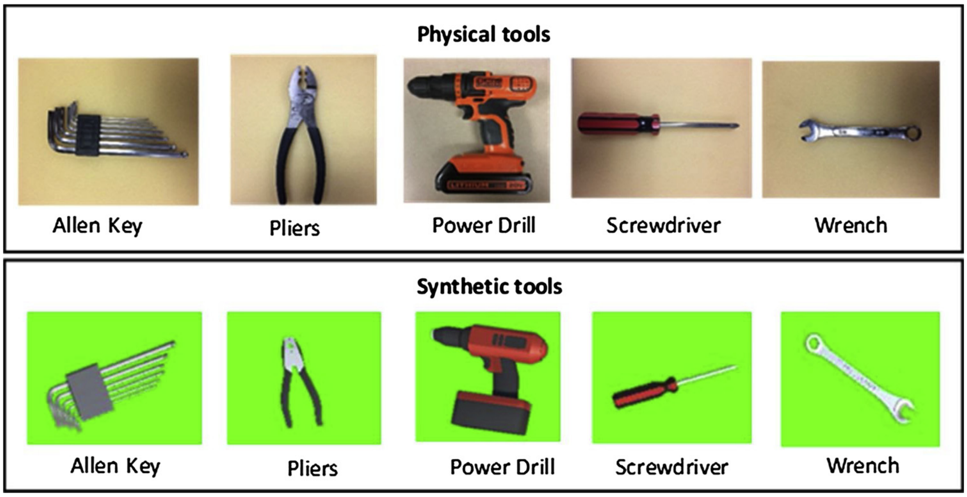


Fig. 10. Comparison between 5 classes of real (upper row) and synthetic (lower row) tools.

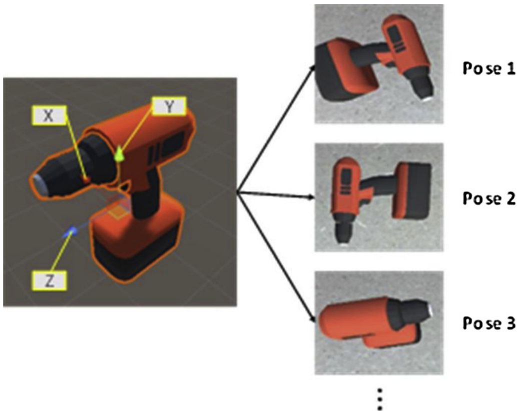


Fig. 11. A synthetic tool rotates about three different axes. To create inner-class variations for the purpose of higher recognition rate, a spatially varying generator is scripted for each class of synthetic tools.

of the dataset is fed into a deep convolutional neural network for feature extraction. Then a Region Proposal Network is implemented to compute high-quality region proposals and pass forward to classification and bounding box regression. The model is trained until the weights are fine-tuned, which minimizes the loss (cost) of the objective function. The objective function is given below:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{4}$$

The two terms on the right side represent losses of classification and bounding-box regression. Symbol i indicates the index of proposals generated while using RPN. p_i^* is the binary classification label which returns 1 when the object in the proposal is detected, and 0 otherwise. The bounding-box regression loss is activated only when $p_i^* = 1$, which contains t_i and t_i^*, attributing to four parameterized coordinates of the predicted box and the ground-truth box.

2.2.2. Synthetic tool dataset

Considering the expensive cost of collecting real tool images and labelling the instances manually, a synthetic tool dataset built with Computer-Aided Design (CAD) models [38] is introduced. The objective is to detect real tools in an assembly scenario by using only computer-generated images for the dataset, which is efficient when training a new classifier and the amount of real data is limited [30,31]. The generation of synthetic dataset for each class of tools is initiated by creating their CAD models with corresponding color, shapes, and textures. Then, the models are transformed to OBJ file format and imported to Unity3D engine for image generation. Each image of the synthetic dataset for five categories of CAD tools (Allen key, pliers, power drill, screwdriver, and wrench) is collected by placing a 3D model onto a 2D image background and project it as a new image with 1024 × 600 pixels. Fig. 10 is the comparison between 5 classes of real and synthetic tools, which shows that the synthetic tools have high similarity with the physical tools. Therefore, they can be used for training deep learning models.

To determine the background for synthetic data generation and number of images, the approach [31] tested on a real PASCAL

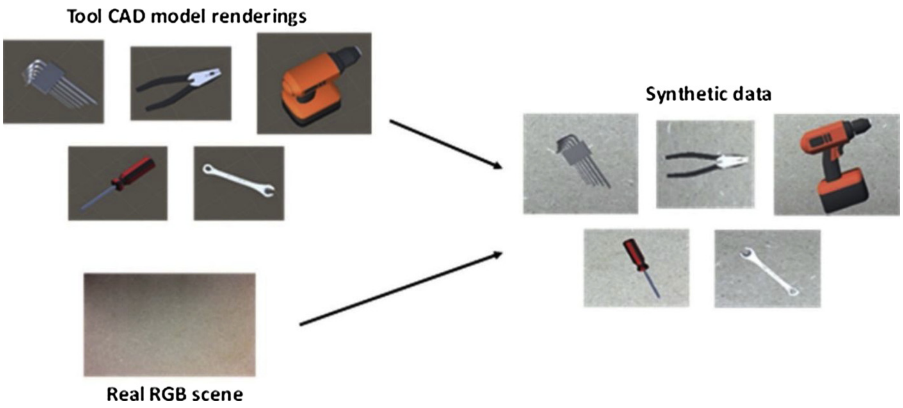


Fig. 12. Synthetic data for R-CNN. By using data augmentation, the synthetic dataset can be generated with CAD models and an RGB scene (background).

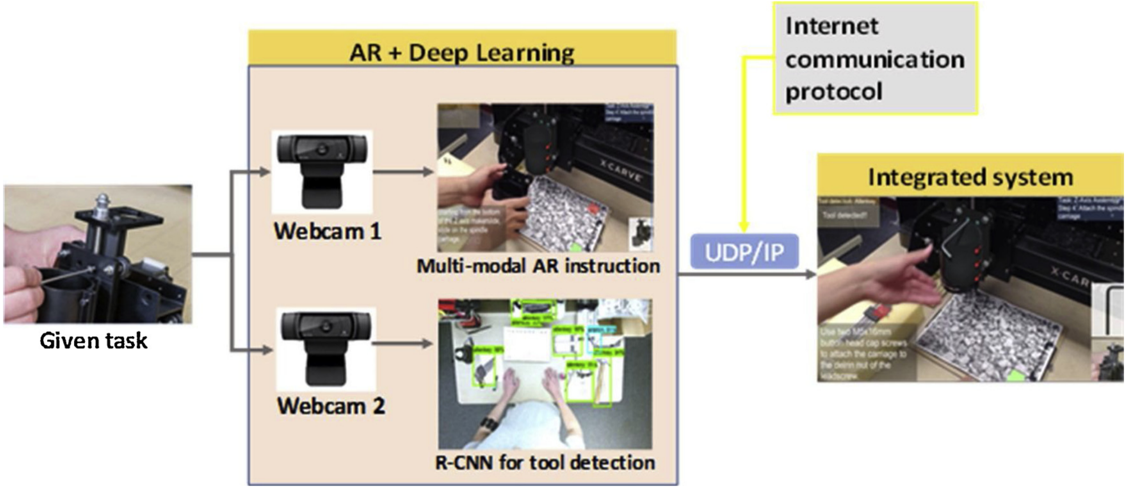


Fig. 13. The workflow of the integrated system.

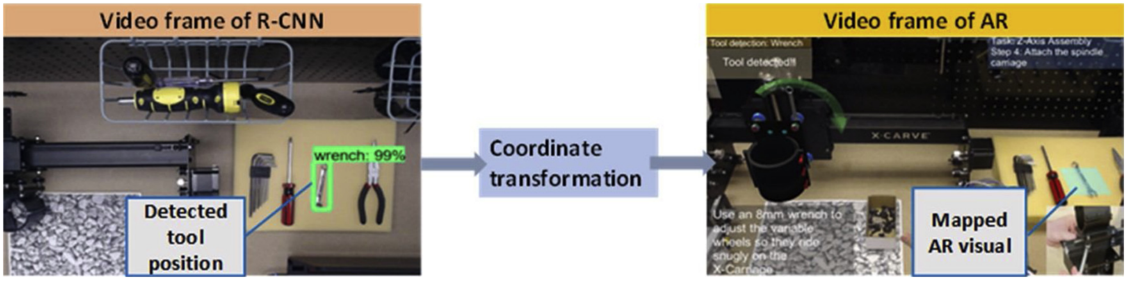

Video frame of R-CNN
wrench: 99%
Detected tool position
Coordinate transformation
Video frame of AR
Tool detection: Wrench
Tool detected!!
Task: Z-Axis Assembly
Step 4: Attach the spindle carriage
Use an 8mm wrench to adjust the variable wheels so they ride snugly on the X-Carriage
Mapped AR visual

Fig. 14. The derived transformation matrix maps the coordinates to a video frame of AR.

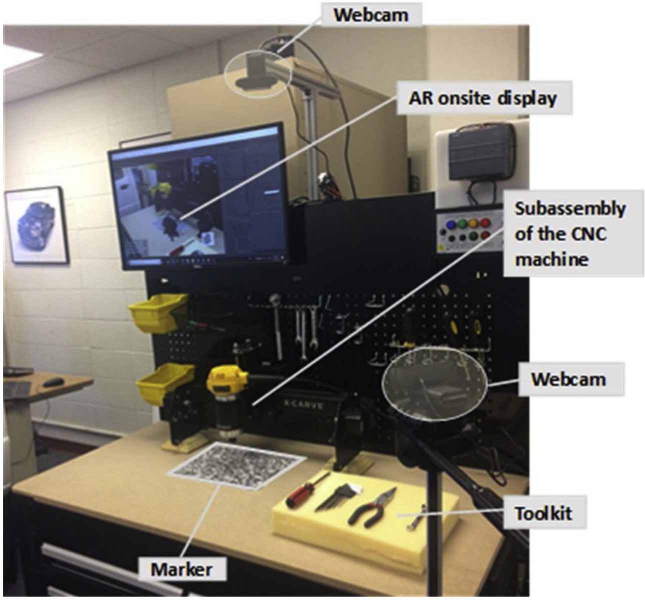


Fig. 15. A workstation setup for the experiment.

VOC2007 dataset [33] with the top mean Average Precision (mAP) having the configuration of RGB model and RGB background is adopted. Also, the number of approximately 2000 images for each training class is defined for the peak result [31]. To create an inter-class variation for reaching higher performance in recognizing objects with different orientations [30,31], a spatially varying generator is scripted and applied to the models, which enables tool models to constantly alter poses while projecting for synthetic image generation. Fig. 11 shows a CAD tool model that rotates about three different axes. Fig. 12 presents the generation of synthetic dataset for each class of tools using data augmentation.

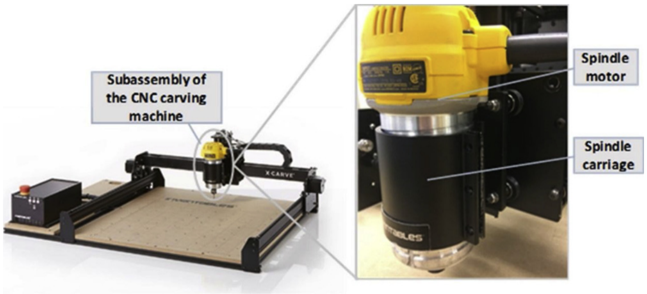


Fig. 16. The installed spindle motor with the subassembly of the CNC carving machine.

2.3. System integration

In order to develop the integrated AR instruction system, two proposed components are connected through a User Datagram Protocol (UDP) socket. For AR tooling message, the detection results of tools including class and coordinates are transmitted through a scripted internet protocol (IP), augmenting computer-generated visuals on the corresponding tools. Fig. 13 presents the workflow of the integrated two-stream system. Two RGB webcams are responsible for AR rendering and tool detection, respectively.

Since the integrated system runs with two webcams, mapping the 2D coordinates from R-CNN to AR for augmenting visual guidance is required. Therefore, an affine transformation is utilized for coordinate transformation. The function of the transformation is:

Table 1
The spindle assembly task.

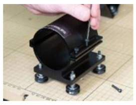
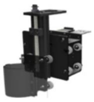
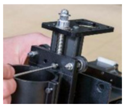
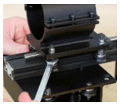
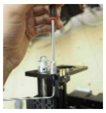
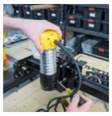
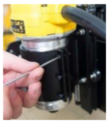
Step No.	Step name	Graphic instruction	Tool/Component	Text instruction
1	Insert spindle carriage clamping bolts		Tool: Allen key Component: Socket head screw M4 x 16 mm	Thread in three of the M4 x 16 mm socket head screws.
2	Attach spindle Carriage to Z-axis (1)		None	Starting from the bottom of the Z axis, slide on the spindle carriage.
3	Attach spindle Carriage to Z-axis (2)		Tool: Allen key Component: Button head cap screw M5 x 16 mm	Use two M5 x 16 mm button head cap screws to attach the carriage to the delrin nut of the leadscrew.
4	Attach spindle Carriage to Z-axis (3)		Tool: 8 mm wrench Component: Lock nut	Use an 8 mm wrench to adjust the variable wheels so they ride snugly on the carriage, but not so tight that they cannot be moved by hand.
5	Attach Z-axis home switch		Tool: Allen key & pliers Component: screw, nut	Thread on an M3 nylock nut and tighten it against the plate with either a 5.5 mm socket as pictured, or with an appropriate wrench/plier.
6	Mount spindle (1)		Tool: Screwdriver Component: Spindle	Use a screwdriver or other prying tool to gently pry the spindle mount open and insert the router until the yellow button shows through on the bottom of the spindle mount.
7	Mount spindle (2)		Tool: Allen key Component: Socket head screw M4 x 16 mm	Tighten the three M4 x 16 mm screws to hold the router in place.

Table 2
Three types of assembly errors.

No.	Error type	Description
1	Tool/Part selection	Misuse the tool/part to conduct the assembly
2	Assembly order	Assemble with incorrect sequence
3	Installation	Assemble with incorrect fixation

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} a1 & a2 & a3 \\ a4 & a5 & a6 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ w \end{bmatrix} \tag{10}$$

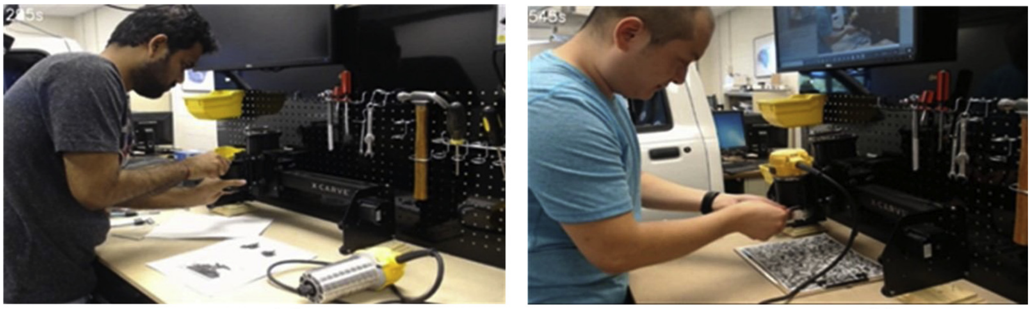
where (x, y, w) represents each center point of three drawn bounding boxes in the image plane of R-CNN. (x', y', w') indicates the three corresponding points in a video frame of AR. $a1...a6$ are the target parameters of the transformation matrix. Once the transformation matrix is obtained from two different sets of points, the transformation can be achieved with the derived matrix. Fig. 14 shows the workflow of mapping an AR visual with coordinate transformation.

3. Experiments

In this section, a designed experiment is intended to evaluate the integrated system's performance, which is a spindle motor installation for a desktop CNC carving machine. The three following subsections detail the setup, evaluation metrics, and subject selection of the experiment.

3.1. Experimental setup

To evaluate the performance of the integrated system, an experimental setup with two Logitech C920 Pro webcams and one monitor for the on-site display are prepared. One webcam is installed on top of the workbench to capture the working area for tool detection, and the second webcam is mounted on a tripod capturing video streams for the AR display. For the assembly, a toolkit is arranged aside. As the AR and



(a) (b)

Fig. 17. Two different subjects perform the experiment with the two different methods: (a) paper manual and (b) smart AR.

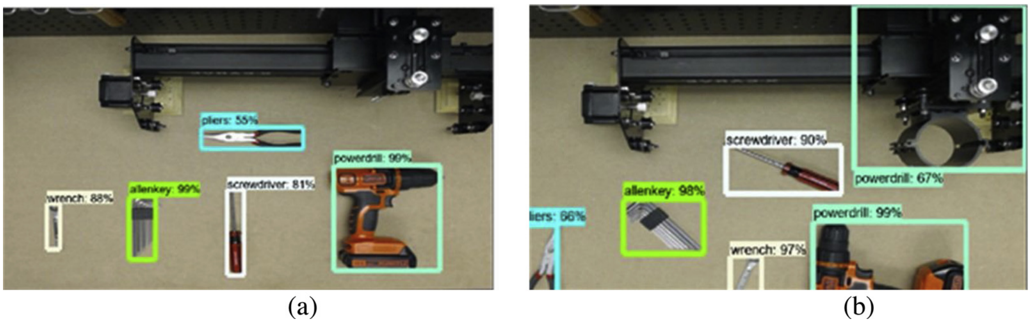


Fig. 18. Tool detection with Faster R-CNN runs with a webcam. The tool detector recognizes tools with various orientations in two different scenarios.

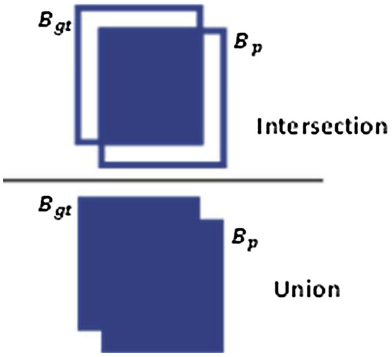


Fig. 19. Illustration of the Intersection over Union (IoU).

Table 3
Average Precision on detecting different real tools.

Tool	Average Precision
Allen Key	64.7 %
Plier	95.9 %
Power drill	72.4 %
Screwdriver	97.2 %
Wrench	93.5 %
mean	84.7 %

object detection processes are running in real-time with a NVIDIA GTX 1080 Ti GPU, the tasks' execution time is not affected. The full workstation setup for the designed experiment is shown in Fig. 15.

The experiment is on the spindle installation of a CNC carving machine (Inventables: X-Carve 750 mm), for which the assembly worker would need instructional guidance to finish the assembly correctly. The goal of the assembly task is to install the spindle motor onto the z-axis mechanism with the components and tools provided. A photo including the subassembly of the CNC carving machine and the spindle motor is shown in Fig. 16.

The spindle motor installation contains seven steps. Each step consists of multiple operations that require different tools or components. A summary of information provided by the instructional manual along with graphical illustrations is given in Table 1 [36]. With the defined assembly task, the experimental setup is determined.

3.2. Subject selection

On subject selection, 20 physically and cognitively healthy subjects, including both male and female, at the average age of 28 at Missouri S& T are recruited. All the subjects are confirmed having no prior knowledge of the experiment and divided into two groups of 10 subjects conducting the experiment. One group is asked to perform the assembly by referring to the paper manual provided by the manufacturer, and the other group performs the assembly using the integrated AR system. During the experiment, each of the subjects is asked to stand in front of the workbench and complete the assembly task.

3.3. Evaluation metrics

This section describes the testing procedure and the evaluation metrics in assessing the performance of the integrated system. In terms of productivity, the completion time and number of errors are the most crucial indicators of the performance in an assembly operation [37]. Thus, two kinds of data are recorded when the experiment is being performed. The elapsed time is recorded with a stopwatch, while the number and types of assembly errors are documented if a mistake occurs. Three types of assembly errors are listed in Table 2, which are considered the most generic errors in a manual mechanical assembly process [37].

As shown in Table 2, the tool/part selection error occurs when a subject misuses a tool/part. The assembly order error is associated with mistakenly assembling a component with incorrect orders. The installation error takes place when a subject installs parts with wrong fixation, which includes mismatching components and securing them improperly.

4. Results and discussion

This section discusses the experimental results including the tool detector and AR rendering, as well as the evaluation of the integrated system. Fig. 17 shows a snapshot each of two subjects performing the

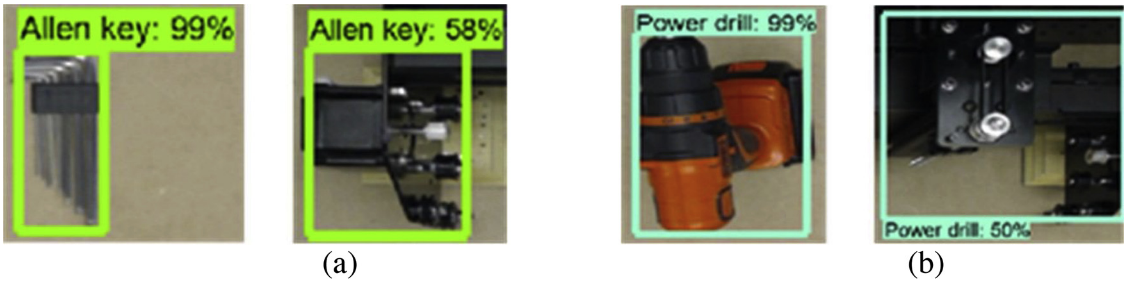
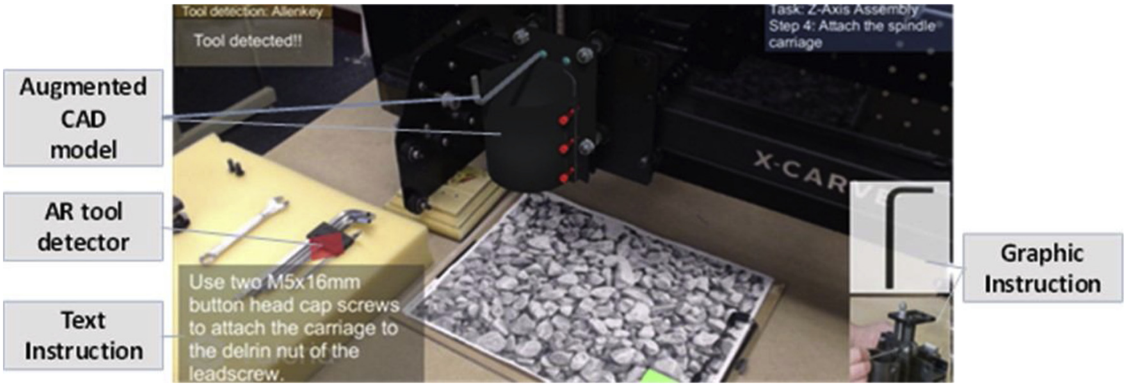


Fig. 20. Two example frames of False Positive (FP). The misclassification occurs when video frames include non-tool objects in the background.



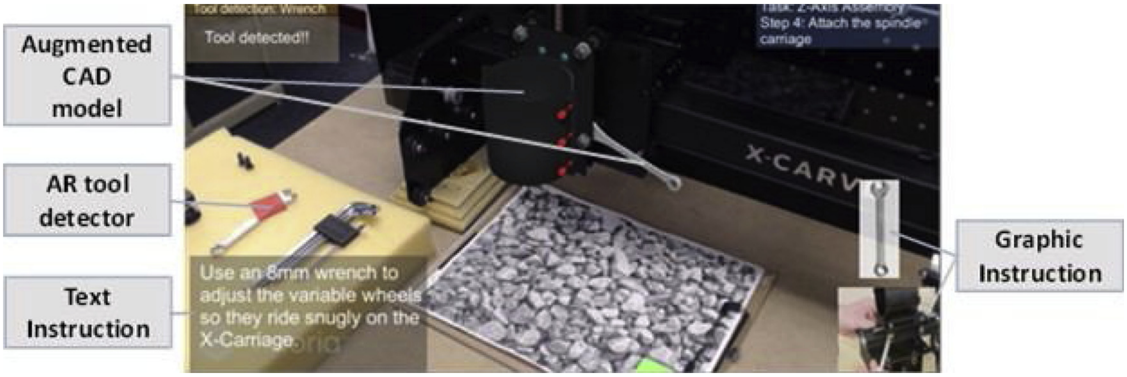
Tool detection: Allenkey

Tool detected!!

Task: Z-Axis Assembly
Step 4: Attach the spindle carriage

Use two M5x16mm button head cap screws to attach the carriage to the delrin nut of the leadscrew.

Fig. 21. AR integrated with the tool detector for Step 3 of the assembly task.



Tool detection: Wrench

Tool detected!!

Task: Z-Axis Assembly
Step 4: Attach the spindle carriage

Use an 8mm wrench to adjust the variable wheels so they ride snugly on the X-Carriage.

Fig. 22. AR integrated with the tool detector for Step 4 of the assembly task.

Table 4
Evaluation results of the group using paper manual.

Subject	Gender	Total number of errors	Number of error type 1	Number of error type 2	Number of error type 3	Completion time (s)
1	Male	6	1	1	4	725
2	Male	2	0	0	2	616
3	Female	3	0	1	2	729
4	Male	2	0	0	2	596
5	Male	4	2	0	2	1057
6	Male	6	1	0	5	712
7	Male	2	1	0	1	689
8	Male	3	3	0	0	605
9	Male	3	1	0	2	708
10	Female	3	2	0	1	1113
Mean	—	34	—	—	—	755

Table 5
Evaluation results of the group using the smart AR instruction system.

Subject	Gender	Total number of errors	Number of error type 1	Number of error type 2	Number of error type 3	Completion time (s)
11	Male	1	0	0	1	424
12	Male	2	0	0	2	359
13	Male	2	0	0	2	531
14	Male	1	0	0	1	531
15	Male	5	1	0	4	600
16	Female	4	0	0	4	914
17	Male	2	1	0	1	421
18	Male	1	0	0	1	573
19	Male	3	0	0	3	413
20	Female	2	1	0	1	278
Mean	—	23	—	—	—	504

experiment using the two different methods (paper manual vs. smart AR) for the system evaluation.

4.1. Performance of the tool detector for AR rendering

The tool detector is developed by fine-tuning a pre-trained Faster R-CNN model using TensorFlow object detection API with approximately 64 K iterations and 3 × 10^{-4} learning rate. The classification layer of the Faster R-CNN algorithm is modified to output probability scores in [0,[1]] over five classes of tools. Once the tools are detected in a video frame, the detector draws bounding boxes around it. Fig. 18 shows two example results of the detected tools.

For the quantitative evaluation of the tool detector, the Intersection over Union (IoU) metric [33] is adopted. If a ground-truth box and a predicted box are overlapped by 50 % or larger, then the prediction is a True Positive (TP). The formula is denoted as:

$$a_o = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{11}$$

where a_o represents the overlap ratio between the ground-truth box B_{gt} and the predicted box B_p; $B_p \cap B_{gt}$ and $B_p \cup B_{gt}$ are the intersection and union of them, respectively. Fig. 19 provides an illustration of IoU.

To calculate precision, the metric [33] is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

where True Positive (TP) represents an instance from the target class that is correctly classified as the target class. False Positive (FP) represents an instance from a class other than the target class that is misclassified as the target class. The tool detection results show that the tool detector can classify and localize the physical location of a real tools with different poses, which demonstrates the viability of CNN using CAD data augmentation. Table 3 shows the average precision of

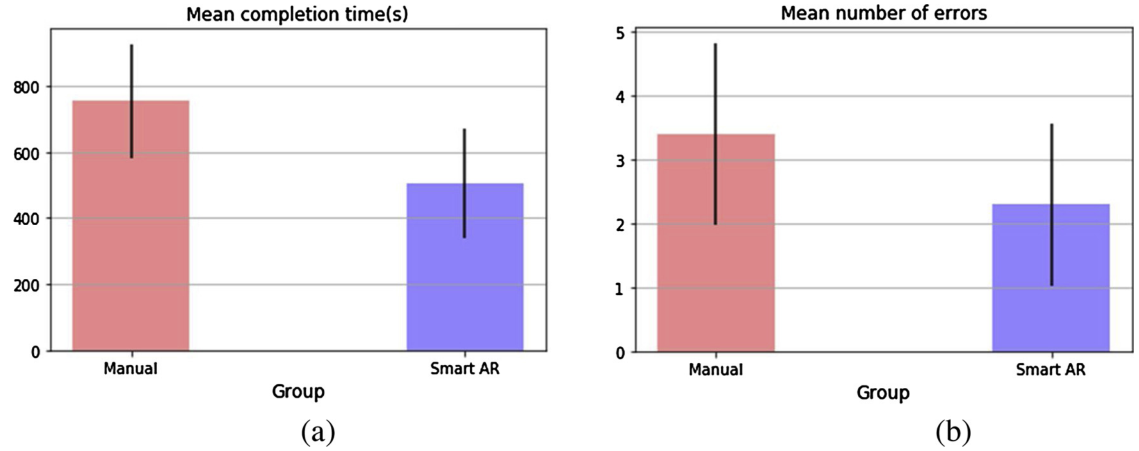


Fig. 23. The mean completion time and number of errors of the paper manual vs. smart AR group.

Table 6
Percentage reduction for each type of errors using the AR instruction.

Error type	1	2	3
Reduction	72.7 %	100 %	4.8 %

Table 7
Error analysis of the experiment with the paper manual.

Error Step	1	2	3	Description
1	3	1	8	Type 1: Should use an Allen key, instead of a screwdriver Type 2: Incorrect assembly sequence Type 3: Should leave the screws loose
2	0	0	12	Type 3: Mismatch the carriage and the track
3	2	1	0	Type 1: Should use an Allen key, instead of a screwdriver Type 2: Should tighten the screws first
4	0	0	1	Type 3: Securing incorrect nuts
5	0	0	1	Type 3: Apply too much torque while securing the screw
6	1	0	0	Type 1: Should use an Allen key, instead of a screwdriver
7	4	0	0	Type 1: Should use an Allen key, instead of a screwdriver

Table 8
Error analysis of the experiment with the smart AR system.

Error Step	1	2	3	Description
1	0	0	3	Type 3: Should leave the screws loose
2	0	0	12	Type 3: Mismatch the carriage and the track
3	2	0	2	Type 1: Should use an Allen key, instead of a screwdriver Type 3: Mismatch the carriage and the delrin nut
4	0	0	0	—
5	1	0	0	Type 1: Should use a plier, instead of a wrench
6	0	0	3	Type 3: Pry at an incorrect position
7	0	0	0	—

detecting five real tools using the tool detector for the Intersection over Union (IoU) evaluation on a real tool dataset captured from the webcam with a resolution of 1024 × 600.

The mean of the Average Precision for the five tools is 84.7 %, indicating a strong performance of utilizing synthetic data for real object detection. As shown in Table 3, the screwdriver outperforms all the other tools in the tool detection, which is likely due to the unique tool shape and color of the grip. Allen key has the lowest score of precision, because its shape and color may result in a confusion with other non-tool objects in the background. Fig. 20 shows two example frames of False Positive (FP). Two irrelevant objects in the bounding boxes are misclassified as Allen key and power-drill with the detection scores of 58 % and 50 %, respectively, indicating a confusion of predicted classes inside the bounding boxes. This indicates that the background affects the precision of tool recognition. Also, a decrease in the precision occurs when there are more objects within the captured frame.

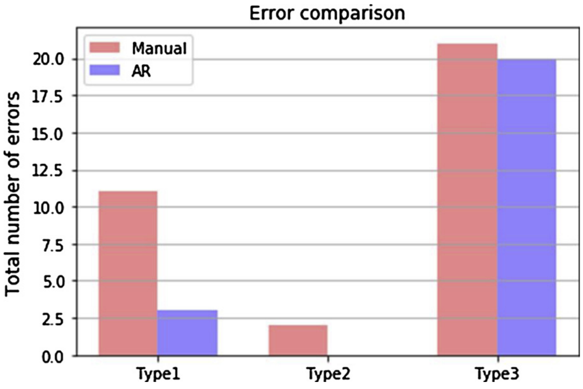


Fig. 24. The comparison of three types of errors.

By integrating multi-modal AR rendering with the tool detector, AR tooling messages are provided. In Figs. 21 and 22, two example frames representing two instances of the integrated system in the assembly are shown. The red rectangle in the figure highlights the position of the tool generated by the tool detector. The AR assembly instructions are also rendered in each frame, displaying a current state of the operation.

4.2. Evaluation of the smart AR instruction system

To assess the integrated AR system, the paper manual provided by the CNC carving machine vendor vs. the AR instruction are compared, and the results are given in Tables 4 and 5. Also, Fig. 23 compares the mean completion time and number of errors of the two groups using ANOVA. By following the smart AR instruction, the completion time and the number of assembly errors are reduced by 33.2 % and 32.4 %, respectively, compared to the paper manual. Fig. 23 and Table 6 present a comparison in the reduction of each type of errors with using the smart AR instruction system.

As shown in Table 6, two types of errors including tool/part selection (Type 1) and assembly sequential order (Type 2) are reduced by 72.7 % and 100 %, respectively, with the aid of the dynamic AR visual support and tool detection. The installation error (Type 3) is recorded as the most error-prone from both groups, which also has the least improvement in error reduction with the assistance of smart AR. Tables 7 and 8 present a detailed summary of errors from the two different groups, which concludes the error type and how the errors are made in each step of the experiment. According to the recorded data, as shown in Fig. 24, Type 3 error in Step 2 is the most recorded type of error using either the paper manual or the smart AR system. Examining these errors provides the following insight: Although AR rendering is able to provide the spatial information and geometry of each part to be assembled, improvement to the AR system is still needed in order to assist the workers on understanding the relations among different parts, e.g., matching the carriage's V-wheel mechanism to the z-axis track. The proposed AR instruction system can be further improved based on the errors observed in the experiment.

To sum up, the experimental results on the manual assembly task indicate a considerable improvement in the assembly performance by introducing smart multi-modal AR instructions to manual assembly tasks, compared to the conventional method of using paper manuals. In addition, the developed system has demonstrated the promising potential of implementing AR with deep learning to assist manual assembly.

5. Conclusion

This paper presents the development of a worker-centered, AR-based mechanical assembly instruction system aimed at improving workers' performance by introducing deep learning to augmented reality (AR) for intelligent manufacturing. The developed system consists of multi-modal AR instructions and a tool detector. The multi-modal AR rendering, which provides various on-site instructions (texts, videos, 3D animations) is realized with homography transformation. The tool detector is developed using a Faster R-CNN model trained on a CAD-based synthetic tool dataset, which detects real physical tools with an average precision of 84.7 %. Experimenting with this system on the spindle motor assembly shows that the system reduces the assembly completion time and number of errors by 33.2 % and 32.4 %, respectively. Thus, the proposed AR system has demonstrated its ability in assisting human operators to perform complex assembly tasks.

Declaration of Competing Interest

I declare no conflict of interest.

Acknowledgments

This research work is supported by the National Science Foundation grant CMMI-1646162 and by the Intelligent Systems Center at Missouri University of Science and Technology. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Byerly J. How augmented reality is revolutionizing job training. Honeywell.com https://www.honeywell.com/newsroom/news/2018/02/how-ar-and-vr-are revolutionizing-job-training (accessed February 20, 2018). 2020.
[3] Kloberdanz K. "Smart specs: OK glass, fix this jet engine." GE.cOm. 2020https://www.ge.com/reports/smart-specs-ok-glass-fix-jet-engine/.
[4] Hu L, Nguyen NT, Tao W, Leu MC, Liu XF, Shahriar MR, et al. Modeling of cloud-based digital twins for smart manufacturing with MT connect. Procedia Manuf 2018;26:1193–203.
[5] Leu MC, Tao W, Niu Q, Chi X. "Virtual bone surgery." bio-materials and prototyping applications in medicine. second edition 2018.
[6] Tao W, Lai ZH, Leu MC. "Manufacturing assembly simulations in virtual and augmented reality." augmented, virtual, and mixed reality applications in advanced manufacturing. 2019.
[7] Caudell TP, Mizell DW. Augmented reality: an application of heads-up display technology to manual manufacturing processes. Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences, 2. 1992. p. 659–69. January.
[8] Azuma RT. A survey of augmented reality. Presence Teleoperators Virtual Environ 1997;6(4):355–85.
[9] Tang A, Owen C, Biocca F, Mou W. Comparative effectiveness of augmented reality in object assembly April Proceedings of the SIGCHI Conference on Human Factors in Computing Systems2003. p. 73–80.
[10] Khuong BM, Kiyokawa K, Miller A, La Viola JJ, Mashita T, Takemura H. The effectiveness of an AR-based context-aware assembly support system in object assembly March 2014 IEEE Virtual Reality (VR)2014. p. 57–62.
[11] Syberfeldt A, Danielsson O, Holm M, Wang L. Visual assembling guidance using augmented reality. Procedia Manuf 2015;1:98–109.
[12] Sanna A, Manuri F, Lamberti F, Paravati G, Pezzolla P. Using handheld devices to support augmented reality-based maintenance and assembly tasks January 2015 IEEE International Conference on Consumer Electronics (ICCE)2015. p. 178–9.
[13] Dalle Mura M, Dini G, Failli F. An integrated environment based on augmented reality and sensing device for manual assembly workstations. Procedia CIRP 2016;41:340–5.
[14] Webel S, Bockholt U, Engelke T, Gavish N, Olbrich M, Preusche C. An augmented reality training platform for assembly and maintenance skills. Rob Auton Syst 2013;61(4):398–403.
[15] Leu MC, ElMaraghy HA, Nee AY, Ong SK, Lanzetta M, Putz M, et al. CAD model based virtual assembly simulation, planning and training. CIRP Annals 2013;62(2):799–822.
[16] Wu S, Tao W, Leu MC, Long S. Engine sound simulation and generation in driving simulator. IISE Annual Conference and Expo 2018:611–6.
[17] Werrlich S, Eichstetter E, Nitsche K, Notni G. An overview of evaluations using augmented reality for assembly training tasks. Int J Comput Inf Eng 2017;11(10):1068–74.
[18] Werrlich S, Nitsche K, Notni G. Demand analysis for an augmented reality based assembly training June Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments2017. p. 416–22.
[19] Tao W, Leu MC, Yin Z. American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion. Eng Appl Artif Intell 2018;76:202–13.
[20] Tao W, Lai ZH, Leu MC, Yin Z. American sign language alphabet recognition using leap motion controller January Proceedings of the 2018 Institute of Industrial and Systems Engineers Annual Conference2018.
[21] Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL. A public domain dataset for human activity recognition using smartphones April Esann.; 2013.
[22] Anguita D, Ghio A, Oneto L, Llanas Parra FX, Reyes Ortiz JL. Energy efficient smartphone-based activity recognition using fixed-point arithmetic. J Univers Comput Sci 2013;19(9):1295–314.
[23] Ward JA, Lukowicz P, Troster G, Starner TE. Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE Trans Pattern Anal Mach Intell 2006;28(10):1553–67.
[24] Tao W, Lai ZH, Leu MC, Yin Z. Worker activity recognition in smart manufacturing using IMU and sEMG signals with convolutional neural networks. Procedia Manuf 2018;26:1159–66.
[25] Al-Amin M, Qin R, Tao W, Leu MC. Sensor data based models for workforce management in smart manufacturing January Proceedings of the 2018 Institute of Industrial and Systems Engineers Annual Conference2018.
[26] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436.
[27] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012:1097–105.
[28] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014:580–7.
[29] Girshick R. Fast r-cnn. Proceedings of the IEEE International Conference on Computer Vision 2015:1440–8.
[30] Hattori H, Naresh Boddeti V, Kitani KM, Kanade T. Learning scene-specific pedestrian detectors without real data. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015:3819–27.
[31] Peng X, Sun B, Ali K, Saenko K. Learning deep object detectors from 3d models. Proceedings of the IEEE International Conference on Computer Vision 2015:1278–86.
[32] Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 2015:91–9.
[33] Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. Int J Comput Vis 2010;88(2):303–38.
[34] Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW. Selective search for object recognition. Int J Comput Vis 2013;104(2):154–71.
[36] Inventables. "X-Carve Instructions." Inventables.com. http://x-carve-instructions.inventables.com/750mm/.
[37] Hou L, Wang X, Bernold L, Love PE. Using animated augmented reality to cognitively guide assembly. J Comput Civ Eng 2013;27(5):439–51.
[38] Leu MC, Tao W, Ghazanfari A, Kolan K. NX 12 for engineering design. Missouri University of Science and Technology; 2017.
[39] Tao W, Lai ZH, Leu MC, Yin Z, Qin R. A self-aware and active-guiding training & assistant system for worker-centered intelligent manufacturing. Manuf Lett 2019;21:45–9.
[40] Zhang Y, Kwok TH. Design and interaction interface using augmented reality for smart manufacturing. Procedia Manuf 2018;26:1278–86.
[41] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition vol. 2. In ICML deep learning workshop; 2015.
[42] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation October International Conference on Medical Image Computing and Computer-Assisted Intervention2015. p. 234–41.
[43] Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 2014;22(10):1533–45.
[44] Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural

machine translation. arXiv preprint arXiv 2015. 1508.04025.

[45] Zhu Y, Mottaghi R, Kolve E, Lim JJ, Gupta A, Fei-Fei L, et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning May 2017 IEEE International Conference on Robotics and Automation (ICRA)2017. p. 3357–64.

[46] Li G, Yang Y, Qu X. Deep learning approaches on pedestrian detection in hazy weather. IEEE Transactions on Industrial Electronics; 2019.

[47] Li G, Li SE, Zou R, Liao Y, Cheng B. Detection of road traffic participants using cost-effective arrayed ultrasonic sensors in low-speed traffic situations. Mech Syst Signal Process 2019;132:535–45.

[48] Li G, Wang Y, Zhu F, Sui X, Wang N, Qu X, et al. Drivers' visual scanning behavior at signalized and unsignalized intersections: a naturalistic driving study in China. J Safety Res 2019;71:219–29.

[49] Esmaeilian B, Behdad S, Wang B. The evolution and future of manufacturing: a review. J Manuf Syst 2016;39:79–100.

[50] Wang J, Ma Y, Zhang L, Gao RX, Wu D. Deep learning for smart manufacturing: methods and applications. J Manuf Syst 2018;48:144–56.

[51] Wang L. From intelligence science to intelligent manufacturing. Engineering 2019;5(4):615–8.

[52] Li P, Jia X, Feng J, Zhu F, Miller M, Chen LY, et al. A novel scalable method for machine degradation assessment using deep convolutional neural network. Measurement 2020;151:107106.