Article

An Intelligent Manufacturing Approach Based on a Novel Deep Learning Method for Automatic Machine and Working Status Recognition

Feiyu Jia 1, Ali Jebelli 1, Yongsheng Ma 2 and Rafiq Ahmad 1,∗

1 Laboratory of Intelligent Manufacturing, Design and Automation (LIMDA), Department of Mechanical Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada; feiyu1@ualberta.ca (F.J.); jebelli@ualberta.ca (A.J.)
2 Department of Mechanical and Energy Engineering, Southern University of Science and Technology, Shenzhen 518055, China; mays@sustech.edu.cn
∗ Correspondence: rafiq.ahmad@ualberta.ca

Abstract: Smart manufacturing uses robots and artificial intelligence techniques to minimize human interventions in manufacturing activities. Inspection of the machine’s working status is critical in manufacturing processes, ensuring that machines work correctly without any collisions and interruptions, e.g., in lights-out manufacturing. However, the current method heavily relies on workers onsite or remotely through the Internet. The existing approaches also include a hard-wired robot working with a computer numerical control (CNC) machine, and the instructions are followed through a pre-program path. Currently, there is no autonomous machine tending application that can detect and act upon the operational status of a CNC machine. This study proposes a deep learning-based method for the CNC machine detection and working status recognition through an independent robot system without human intervention. It is noted that there is often more than one machine working in a representative industrial environment. Therefore, the SiameseRPN method is developed to recognize and locate a specific machine from a group of machines. A deep learning-based text recognition method is designed to identify the working status from the human–machine interface (HMI) display.

Keywords: smart manufacturing; object detection; text recognition; Siamese neural network

1. Introduction

With current labor shortages and high costs, workers’ health, safety, and product quality may be compromised to compensate for manufacturing productivity [1]. Most manufacturing activities, such as machine tending tasks, are low-tech, repetitive, and dull, so workers can easily be replaced. With the development of robots, they are gradually being applied to machine tending tasks by collaborating with operators and even replacing workers to load and unload the raw parts to the machines. Recently, numerous research and development have been carried out and contributed to the robotic system in smart manufacturing [2–7]. Since robots can sense and respond to different scenarios for various manufacturing tasks, collaborative robots (cobots) have been widely implemented to relieve workers from risky, tedious, and repetitive manufacturing tasks [8]. In addition, it ensures high throughput rates and low costs [9].

Collaborative robots work in four different manners according to different collaborative environments [10], which are: (1) Coexistence: the robot and operator are in the same environment but generally do not interact with each other; (2) synchronized: the robot and worker are in the same workspace, but at different times; (3) cooperation: the worker and robot are in the same workspace at the same time, but each focuses on separate tasks; and, (4) collaboration: the robot and operator execute the same task, which means the action of
the one has immediate consequences on the other. However, these four types of cobots still suffer from some challenges and drawbacks. Safety is the primary concern. Although some perception sensors are utilized on cobots to avoid injuries, operators must be careful all the time working with cobots if they move too fast. Moreover, cobots are often dedicated and fixed for a particular machine, and programmed for a specific task. It is an inefficient use of factory space and causes high costs to satisfy all manufacturing activities. In addition, it requires experienced engineers to reprogram the cobots to meet different production requirements when changing the manufacturing tasks.

The autonomous mobile manipulator (AMM) is proposed to extend the capabilities of conventional collaborative robots [11]. The architecture of AMM is a robot arm mounted upon a mobile platform, which combines locomotion capability with manipulation ability. Since it combines collaborative and mobile robot characteristics, it is more flexible and adaptable to changes in tasks or environments. Although AMM can perform more manufacturing tasks than traditional cobots cannot, there is still a need for experienced operators to be onsite to assist the AMM in finding the target machines, checking the machine’s working status, and taking emergency actions in case of problems. For example, the robot does not load raw material properly.

As discussed in [12], smart devices and intelligent solutions can significantly improve the manufacturing process with artificial intelligence and imaging equipment development. They have gradually become a hot topic in smart manufacturing [13]. In [14], authors propose an intelligent perceiving and planning system based on deep learning for a 7-DoF manipulator with a vision system. The vision system and designed intelligent process enhance collaboration ability through recognizing the target objects and improving the efficiency of robot planning. In [15], a collaborative robotics framework for top view surveillance is proposed. This study adopts the pre-trained deep learning models for object detection and localization to assist human operators in managing and controlling different applications. However, there is a challenge to recognize objects because the appearance of objects change significantly due to the change in camera position and shooting angle, which results in weak performance. In [16], authors develop a deep learning-based object detection method for the mobile robot manipulator in small and medium-sized enterprises production. This study applies region convolutional neural networks to recognize and localize the charging station and printing machine to achieve automatic tag production. In addition, it designs a human detection algorithm for the manipulator to increase safety while collaborating with operators.

Most recent studies and existing applications in smart manufacturing adopt mobile manipulators with the vision system and designed object detection methods to improve manufacturing production safety, efficiency, and intelligence. In those studies, humans often take great responsibility to control and operate robots through computer programs. Consequently, imitating the humans’ behavior, such as the “eye-brain-hand” process, is necessary to realize intelligent and autonomous manufacturing production truly and has become a widespread trend [14]. Therefore, this study proposes a deep learning-based intelligent manufacturing approach, aiming to achieve true intelligence and autonomous manufacturing production and machine tending. Compared with other solutions, the benefits of the proposed approach are low costs and high efficiency for small and medium-sized enterprises. In addition, this approach can be applied in lights-out manufacturing to reduce the workload for operators who need to tend machines and tackle the problems on the production floor. In the current study, an autonomous mobile manipulator with a vision system is adopted. The main objectives of this work are to develop an intelligent object detection method for the target CNC machine and HMI display detection in a complex environment and to design a text recognition method for the machine’s working status recognition to further assist in autonomous robot decision-making and problems handling.

The remainder of this paper is organized as follows: in Section 2, an overview of object detection and text recognition is presented. Section 3 explains the methodology of the proposed intelligent manufacturing approach for automatic machine detection and
working status recognition. The case studies and results are demonstrated in Section 4. Section 5 outlines some limitations and potential future work. Finally, conclusions are presented in Section 6.

2. Literature Review

2.1. Object Detection

With the development of computer vision technology, object detection has been applied to many areas, including face detection, pedestrian detection, and traffic sign/light detection. Intelligent manufacturing mainly utilizes object detection techniques, such as quality management, product sorting, packaging, and assembly line. There are two categories of object detection given in recent research: traditional object detection methods and deep learning-based detection methods.

2.1.1. Traditional Object Detection Methods

Most classical object detection methods are usually developed based on low-level and mid-level features, such as color, shapes, edges, and contours. In the late 1990s and early 2000s, there were several milestones in object detection methods dominated by handcrafted features. Scale Invariant Feature Transform (SIFT) is one of them which transforms an image into a wide set of locally scale-invariant features. These scale-invariant features are invariant to image translation, rotation, scaling, illumination, occlusion, and 3D projection. Highly distinctive scale-invariant keypoints are further proposed to match individual features with features from known objects, and then the clusters belonging to the same object are identified through the Hough transform finally verified by least squares. In [17], a local feature descriptor, Histogram of Oriented Gradients (HOG), was first proposed. It counts the occurrences of gradient orientation in localized portions based on a dense grid of uniformly spaced cells and uses overlapping local contrast normalizations to improve accuracy. Despite being similar to SIFT descriptors, it is a significant improvement and gives excellent results for human detection. Because of the advantages of HOG, it has been an important architecture for many object detection methods. As an extension of HOG, the Deformable Part-based Model (DPM) [18] is the best classical object detection method. DPM achieves improvement in both precision and efficiency compared with the previous methods. It considers an object as a global template covering an entire object and a collection of part templates. Then the models are trained discriminatively through support vector machine. However, as the performance of handcrafted features reached its limits, object detection has fallen into a bottleneck period.

2.1.2. Deep Learning-Based Detection Methods

Deep learning-based object detection breaks the deadlocks as deep convolutional neural networks can learn an image’s robust and high-level features. With the development of the GPU computing resources and the availability of large-scale datasets, many deep learning-based object detection methods have been developed. This section introduces the milestone frameworks in object detection because almost all detectors proposed over the last seven years use one of them as the foundation. In general, deep learning-based object detection methods are grouped into two main categories: two-stage methods and one-stage methods.

Regions with Convolutional Neural Network (R-CNN) is the first proposed deep learning-based object detection algorithm. The basic idea is extracting enormous proposals as candidate regions based on selective search [19], and then scaling all proposals to fixed-size images and feed them into a pre-trained CNN model to obtain object features. Finally, using Support Vector Machine (SVM) classifiers to predict the presence of the object and to recognize the object category. However, the redundant feature computations on many overlapped proposals lead to a plodding detection speed. Therefore, Ross Girshick proposed Fast-RCNN [20] to improve R-CNN. It can train a detector and predict
bounding box simultaneously under the same network configurations, which improves the detection speed.

Faster-RCNN [21] is proposed shortly after the Fast-RCNN. Its main contribution is the design of the Region Proposal Network (RPN). RPN can generate region proposals quickly and efficiently. Although Faster-RCNN breaks through the speed bottleneck, computation redundancy occurs at the subsequent detection stage. Joseph Redmon presented a method called You Only Look Once (YOLO) [16] to solve this problem. This algorithm applies a single neural network to simultaneously divide the image into candidate regions and predict probabilities containing the object for each region. Although YOLO improves the detection speed significantly, it suffers from a low accuracy in localization compared with previous methods, especially in some small object detection. Meanwhile, the proposed Single Shot MultiBox Detector (SSD) [17] introduces the multi-reference and multi-resolution detection techniques to improve the detection accuracy while maintaining high detection speed.

Although SSD and YOLO-based methods achieve good accuracy and computation efficiency performance, those models require too much-labeled data for training. However, sometimes much data are not available or easy to obtain for a specific object or task. These methods also suffer from bad performance for small object detection tasks. In addition, those milestone methods mainly focus on classifying and detecting different object categories, so it is not suitable for detecting a particular target CNC machine from many machines in the industrial environment. In this study, Siamese Neural Networks [22] are integrated with the region proposal network to solve this problem. It helps build models with good accuracy and efficiency, even with fewer data and imbalanced class distribution.

2.2. Scene Text Recognition

Scene images contain abundant and precise information, especially the text in the scene images, which is helpful for people to understand the surrounding environment. Scene Text Recognition (STR) is developed as one of the research fields of computer vision to recognize the text from scene images and convert it into machine-readable information [23]. STR consists of two stages, which are text detection and text recognition. Text detection aims to determine whether there is text in a given image or video and to localize the text using bounding boxes. Text recognition aims to identify the detected text and translate the text into machine-readable information [24]. Recently, scene text recognition applications gained much popularity in many fields, such as car plate recognition, product labeling, sorting, and packaging.

Traditional text detection and recognition algorithms often adopt handcrafted features to separate the text and non-text regions in a scene image, requiring demanding and intricate image processing steps [25]. Since traditional methods are not robust, challenging to implement, and constrained by the complexity of creating handcrafted features, they can hardly deal with intricate circumstances. Therefore, this section only discusses recent deep learning-based methods.

2.2.1. Text Detection

Given the similarity between text detection and general object detection, most text detection methods use a general object detection framework as core architecture. Combining a random forest word classifier and convolutional neural network (CNN) was proposed to achieve a high recall rate and precision [26]. In [27], Faster-RCNN was enhanced by the inception region proposal network, responsible for obtaining text candidate regions. Then an iterative bounding box voting scheme was applied to ensure high recall and the best results. In [28], an image processing method named non-maximum suppression was proposed to reduce overlapped and redundant effects. An SSD-based text detection method called TextBoxes ++ [25] was proposed to detect arbitrary-oriented scene text with high accuracy and efficiency using a quadrilateral rectangle without post-processing steps involved, such as non-maximum suppression. However, the accuracy of region proposal-based methods heavily relies on the candidate regions’ generation. Unlike general objects,
the text usually has varying aspect ratios. Therefore, it is necessary to manually design anchors with different aspect ratios and scales, which makes text detection complicated and inefficient.

2.2.2. Text Recognition

Most text recognition methods recognize the scene text by grouping the recognized characters, making text recognition inefficient and difficult for real-time applications. In [29], and accurate scene text recognition method without character-level segmentation was proposed based on a Recurrent Neural Network (RNN). Shi [30] presented a convolutional recurrent neural network (CRNN) to recognize the scene text in arbitrary lengths in scene images by stacking CNN and RNN. This method is not limited to any predefined lexicon and can achieve remarkable performances in scene text recognition tasks. In [31], recursive recurrent neural networks with attention mechanisms [32] were developed to achieve good performance for dictionary-free scene text recognition. Liao [33] proposed a TextBoxes method by combining the SSD and CRNN to speed up the text recognition task to recognize text in scene images.

Although numerous studies have been done, the limitations and challenges of scene text recognition while applying it to real-world manufacturing tasks have not been researched. Efficiency and precision are two main concerns using text recognition methods to obtain the working state of a CNC machine. Text blurring often happens when the mobile robot moves due to camera shake and de-focus, which degrades recognition accuracy [34]. In addition, efficiency is a shortcoming of deep learning-based methods, making it challenging to deploy those methods on mobile devices and lightweight systems [35]. Therefore, this study integrated some pre-processing steps that reduce the complexity of images into the developed text recognition method to improve image quality. In addition, the combination of HOG and CRNN can realize real-time working status recognition of CNC machines while ensuring accuracy.

3. Methodology

There are three main targets of this research. One is to automatically recognize the target CNC machine in a complex environment through a camera mounted on a moving mobile manipulator. Simultaneously, detect the human-machine interface (HMI) of the CNC machine. Once the HMI is detected, it can identify the working status by recognizing the text on the HMI display. This study proposes a novel deep learning-based approach to achieve these goals. First, the Siamese region proposal network (SiameseRPN) method was proposed to achieve the target CNC machine and HMI detection. Then, the detected HMI images are extracted, pre-processed, and used as inputs for working status recognition. Finally, the novel text recognition method by combining the projection-based segmentation and the convolutional recurrent neural network (CRNN) was developed to identify the working status of the target CNC machine. Figure 1 presents the flowchart of the proposed intelligent manufacturing approach. The following section explains the SiameseRPN method and the machine’s working status recognition method.
3.1. Siamese Region Proposal Network (Siameserpn) Architecture

Siamese region proposal network is proposed for the target detection consists of two subnetworks: Siamese neural network and the region proposal network. The framework of SiameseRPN is shown in Figure 2.

3.1.1. Siamese Neural Network

The Siamese Neural Network (SNN) network [22] has been proven effective in the object tracking domain. It uses two branches, which share the same parameters, to learn the similarity between each other. The objective of SNN is to understand the embedding space, which places similar items nearby. In other words, SNN is trained with positive and negative pairs of objects, where positive pairs correspond to samples that need to stay close in the embedding space while negative pairs need to stay far away. Although SNN has drawn significant attention because of its balanced accuracy and speed, it lacks this drawback. In this study, SNN adopts a convolutional neural network without padding to extract the features. Assume the term \( L_T \) represents the translation operator of the kernel; then remove all the paddings to satisfy the full convolution with strike \( k \).

In Equation (1) [22], \( h \) is a fully convolutional function with integer stride \( k \) that maps signals to signals for any translation \( \tau \):

\[
h(L_{k\tau}x) = L_{\tau}h(x) \tag{1}
\]

In this subnetwork, the template frame and the detection frame are fed into the SNN as inputs. The template frame is the previous frame in the video and the detection frame...
is the current frame in the video. In addition, these two branches have the same hyper-parameters and transformations. For convenience, two outputs of the SNN are represented by $M(z)$ and $M(x)$, which are the feature map representations of the template branch and detection branch.

### 3.1.2. Region Proposal Network

The region proposal network (RPN) is first proposed in Faster-RCNN [21], which can extract precise candidate regions quickly and efficiently. Therefore, it makes proposal generation very effective while achieving high accuracy.

In this subnetwork, RPN consists of two parts: pairwise correlation and supervision. The supervision consists of two branches: regression and classification. The regression is to generate the candidate regions, and the classification is to distinguish the foreground and background of input images. If $n$ anchors are automatically generated, RPN needs $2n$ representation channels for the classification task and $4n$ presentation channels for the regression task. Therefore, the pairwise correlation needs to divide $M(z)$ into two datasets $[M(z)]_{\text{cls}}$ and $[M(z)]_{\text{reg}}$, and divide $M(x)$ into $[M(x)]_{\text{cls}}$ and $[M(x)]_{\text{reg}}$. Then, $[M(z)]$ is used as the correlation kernel to compare similarity with $[M(x)]$. Finally, the regression correlation $C_{\text{reg}}$ and classification correlation $C_{\text{cls}}$ are computed through Equation (2) [22]:

$$C_{\text{reg}} = [M(z)]_{\text{reg}} * [M(x)]_{\text{reg}},\quad C_{\text{cls}} = [M(z)]_{\text{cls}} * [M(x)]_{\text{cls}}$$

where $[M(z)]_{\text{reg}}$ and $[M(z)]_{\text{cls}}$ are used as convolutional kernels, and the sign * is the convolution operation.

Smooth $L_1$ loss function with normalized coordinates and cross-entropy loss function are used for regression and classification, respectively. If $P_x$, $P_y$, $P_w$, $P_h$ represent the coordinates of the center point, the width and the height of predicted bounding boxes, and $T_x$, $T_y$, $T_w$, $T_h$ represent those parameters of the ground truth. Then normalized distances between the prediction and ground truth can be computed. RPN is optimized by minimizing the loss function in Equation (3) [22]:

$$\text{Loss} = L_{\text{cls}} + \lambda L_{\text{reg}}$$

where $\lambda$ is the hyper-parameter to balance the performance of regression and classification. $L_{\text{reg}}$ is the loss function for regression, and $L_{\text{cls}}$ is loss function for classification.

### 3.2. Working Status Recognition

The study proposes a novel text recognition method combining HOG and CRNN to accurately and efficiently identify the working status of the CNC machine. This method consists of two stages: text detection using projection-based segmentation [23], and text recognition using convolutional recurrent neural network.

The proposed segmentation algorithm has three steps. First, the projection scans horizontally to determine the upper and lower bounds of the text area. Second, the projection scans vertically to define the left and right bounds of the text area, and then bounding boxes containing the text information are generated. Finally, the non-text regions are removed and the feature maps of text regions are extracted for recognition.

Before applying this segmentation algorithm, some pre-processing steps are adopted. First, transform the color images into grayscale images. The pixel value is from 0 to 255 in a grayscale image. Second, the binarization will convert the grayscale images to binary images using thresholding. Pixels with a greater value than the threshold $P$ are replaced with white color and other pixels are replaced with black color. Here the value of the black pixel is set to one, and the value of the white pixel is set to zero. That is:

$$\begin{align*}
\text{Value}(i, j) &= 1, \quad \text{pixel}(i, j) < P \\
\text{Value}(i, j) &= 1, \quad \text{pixel}(i, j) > P \\
(0 < i < \text{width}, 0 < j < \text{height})
\end{align*}$$

(3)
Width and height are two image parameters, and \( m \) is a variable threshold according to the image background color. The number of black pixels accounts by the Equation:

\[
\text{pixelRow}[j] = \frac{-1}{N \times \text{width}} \sum_{i=-m}^{m+1} \text{Value}(i,j) (0 < j < \text{height}, m = 0, 1, 2, \ldots N - 1) \tag{5}
\]

Here divide the whole image into \( N \) parts with the same height and width. Then, an appropriate threshold \( Q \) is set to separate the text and non-text pixels. In the Equation (6), when \( \text{pixelRow}[j] > Q \), it is set to one, otherwise it is set to zero:

\[
\begin{align*}
\text{Horizontal}[j] &= 1, \text{pixelRow}[j] > Q \\
\text{Horizontal}[j] &= 1, \text{pixelRow}[j] < Q \\
(0 < j < \text{height})
\end{align*}
\tag{6}
\]

\( \text{Horizontal}[j] \) is calculated, traversing the entire array. When the array’s value changes from zero to one, record this pixel as the start point. When the array’s value changes from one to zero, record this pixel as the endpoint. Finally, the text bounding boxes are generated.

After text detection, CRNN [24] is used to recognize the detected text. The framework of CRNN used in this study is shown in Figure 3. It consists of three subnetworks: convolutional layer, recurrent layer, and transcription layer. The convolutional layer automatically extracts feature map sequences from input images. Then the extracted feature sequences are fed into the recurrent network to predict each feature sequence. The transcription layer is used to translate each frame prediction generated in the recurrent layer and group them into a labeled sequence. The main benefit of the CRNN is that it can be trained with one loss function despite being composed of different network architectures.

![Figure 3. The architecture of CRNN.](image-url)
The convolutional layer in CRNN is modified from a standard CNN model by removing the fully connected layers. It extracts feature map representation from an input image. All the input images are scaled to the same size in the pre-processing step. Then feature vector sequence is obtained after the feature maps are extracted in the convolutional layer. To keep the order of the text unchanged, each feature vector is generated from left to right on the feature maps, and the $i$-th feature vector is the concatenation of the $i$-th columns of all the feature maps. Finally, those sequences will be fed into the recurrent layer as inputs.

The feature maps are divided into many rectangle regions with a fixed width, which is the receptive fields. Each feature vector sequence can represent one rectangle region, which is illustrated in Figure 4.

Figure 4. The receptive fields.

Although CNN has proven good performance in recognizing the general objects, it is not suitable for text detection because of its various aspect ratios. Those feature maps extracted from the convolutional layer need to be transformed into sequential representations to avoid being affected by the length variation. Therefore, the recurrent layer modified from the recurrent neural network is adopted to solve this problem. The main advantage of RNN is that it has the stable ability to capture the context in a sequence, which is more useful than identifying each character individually because some ambiguous characters can be easily predicted by observing the contexts.

The transcription layer aims to convert the predicted labeled distribution into a label sequence with the highest conditional probability. Temporal Classification (CTC) [36] is used to calculate the conditional probability. The conditional probability in this study is defined as the label sequence conditioned on the per-frame prediction. The expression of the conditional probability is shown in Equation (7) [24]:

$$p(l|y) = \sum_{\pi:B(\pi)=l} p(\pi|y)$$

where $l$ is the label sequence, $y$ is the predicted labeled distribution obtained from the recurrent layer ($y = y_1, y_2 \ldots y_T$), where $T$ is the length of the input sequence. $B$ is a mapping function, which maps the sequence $\pi$ onto $l$ by removing the duplicate labels and blank labels. For example, $B$ maps the sequence "-m-7-7-" ("-" represents blank label) onto "m77", which is shown in Figure 3.

3.3. Data Augmentation

Although this study adopts enough data to train the proposed model, images of CNC machines tend not to have extensive features due to the limited number of different CNC machines. As a result, increasing the diversity of the training data is widely used to improve the generalization and reduce overfitting [37]. Some geometric distortions are randomly added in this study, including rotation, translation, scaling, vertical flipping, and image distortions, such as Gaussian blur and noise.

3.4. Transfer Learning

Deep learning models often require a large number of input images as training data. However, it is tough to collect enough images for some applications. Transfer learning provides an alternative strategy to address this problem [38] by using a pre-trained deep
learning model as a starting point for another training task rather than building a model from scratch. This study adopted the modified AlexNet pre-trained from ImageNet with the parameters [39], significantly improving training efficiency.

4. Experiments and Results

4.1. Robot System Structure

Figure 5 shows the DOBOT CR5 manipulator used in the robot system. It is fixed on a moveable desk that can move around the working environment. Specifications of this manipulator are demonstrated in Table 1. A webcam with 1920 × 1080 pixels resolution is mounted on the top of the robot arm, used to sense the surrounding environment.

Figure 5. The structure of robot system.

Table 1. Specifications of robot manipulator.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>25 kg</td>
</tr>
<tr>
<td>Maximum Payload</td>
<td>5 kg</td>
</tr>
<tr>
<td>Max Reach</td>
<td>1096 mm</td>
</tr>
<tr>
<td>Voltage</td>
<td>DC48 V</td>
</tr>
<tr>
<td>Maximum Speed of TCP</td>
<td>3 m/s</td>
</tr>
<tr>
<td>Communication</td>
<td>TCP/IP, WIFI</td>
</tr>
<tr>
<td>Power</td>
<td>150 W</td>
</tr>
<tr>
<td>Axes</td>
<td>6</td>
</tr>
</tbody>
</table>

4.2. Training Details

The webcam was used to collect the CNC machine and HMI images. The dataset includes training and testing images with a resolution of 960 × 1080. Images were gathered from different positions in our lab. Two hundred images, including 160 training images and 40 testing images, are used to train and test the developed models. The experiments were conducted on a laptop equipped with an Intel Core i7-8750H 4.0 GHz CPU and a single NVIDIA GeForce GTX 1060 under the Ubuntu 18.04 64-bit operating system.

During the training process, Stochastic Gradient Descent (SGD) is adopted to train the proposed SiameseRPN based on the pre-trained SNN model using ImageNet. In addition, picking positive and negative training samples is also necessary. The criterion used in the target CNC machine detection and HMI detection tasks is based on intersection over union (IoU) and two thresholds \( T_{\text{high}} \) and \( T_{\text{low}} \). When the predicted bounding boxes that have \( \text{IoU} > T_{\text{high}} \) with respect to the ground truth, they are considered positive samples, which means
the correct results. When the $IoU < T_{\text{low}}$, those are considered negative samples and will be removed from the results; the $T_{\text{high}}$ and $T_{\text{low}}$ are set as 0.3 and 0.6 in this study, respectively.

4.3. Evaluation

This section evaluates the proposed methods using the video recorded by the webcam mounted on the robot manipulator in the lab environment. Figure 6 shows the case study for detecting the target CNC machine and HMI display. The robot system moves from the lab door toward the target CNC machine and records the real-time video through the mounted camera. Based on the proposed SiameseRPN method, the target CNC machine and HMI display are recognized and located using green and red bounding boxes, respectively. Figure 7 presents particular examples of recognizing the text information on the HMI display using the proposed working status recognition method, such as the basic instruction information and G-codes. Once identifying the text, it is converted into machine-readable text that assists the autonomous robot system in recognizing the real-time working status of the machine, which is significant for the robot system to achieve further decision-making and execution actions to tackle the emergency and abnormal conditions in a completely autonomous environment.

![Figure 6](image.png)

**Figure 6.** Validation results for the target CNC machine and HMI detection. (a–d) Detect the target CNC machine in the green bounding box, and obstacles such as the table in the yellow bounding box; (e–g) detect the HMI in the red bounding box.
The model’s accuracy for training and testing can reach 66.3% and 61.9%, respectively. The performance is acceptable considering the size of the dataset. Different parameter values, Table 2 compares our proposed method with three milestone methods in object detection: Faster-RCNN, SSD, and YOLO. Two hundred frame images from the recorded video were used for the validation. The accuracy of the proposed method is shown in the equation.

\[ \text{Accuracy} = \frac{N}{T} \]  

where \( N \) is the number of correctly detected objects, and \( T \) represents the total number of images used.

Table 2. Comparison of accuracy with benchmark methods (the target CNC machine detection).

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN</td>
<td>0.67</td>
</tr>
<tr>
<td>SSD</td>
<td>0.54</td>
</tr>
<tr>
<td>YOLO</td>
<td>0.53</td>
</tr>
<tr>
<td>The Proposed Method (this paper)</td>
<td>0.78</td>
</tr>
</tbody>
</table>

For the HMI detection, the accuracy of the proposed method is shown in Figure 9. The model’s accuracy for training and testing can reach 66.3% and 61.9%, respectively. The performance is acceptable considering the size of the dataset. Different parameter values,
such as the batch size of training data, the learning rate of the model, and the activation function, have been applied, but they achieved limited benefit. Therefore, increasing the training dataset size would be better to improve the models’ performance.

One thousand text samples from the HMI display were adopted to validate the proposed working status recognition method. The benchmark method RNN was compared with the proposed method in this study to estimate the performance of the proposed method. The results for both methods are presented in Table 3. It can be concluded that the proposed method improves the recognition performance in precision without the loss of efficiency.

![model accuracy](image)

**Figure 9.** The accuracy of developed model for HMI detection.

**Table 3.** Compared results for RNN and the proposed method (working status recognition).

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN</td>
<td>78.2%</td>
</tr>
<tr>
<td>The Proposed Method (this paper)</td>
<td>85.7%</td>
</tr>
</tbody>
</table>

5. Discussion and Limitations

This paper presented a novel framework to support autonomous machine tending in lights-out manufacturing, which minimizes the involvement of machine operators during the production process. Most machine tending robot systems adopt collaborative robots that connect to the machine via Ethernet to perform various tasks efficiently under the inspection and tending of onsite operators [40–42]. Autonomous mobile robots can use our proposed architecture to achieve fully autonomous machine tending through the integrated vision system and computer vision-based algorithms. The proposed intelligent manufacturing approach can truly achieve autonomous machines, HMI, and emergency detection without the operator’s assistance to support various machine-tending tasks. The proposed methods are flexible, scalable, and adaptable; however, it has some limitations. Text detection and recognition are more sensitive to image quality than general object detection [43]. Sometimes, the mobile camera captures images and videos suffering from poor lighting conditions, such as shadow in the images or reflection of light due to inappropriate shooting distance or angle, making the feature extraction process challenging. Moreover, camera motion or shaking while capturing images can cause text to be blurred [35].

Further research to address the problems mentioned above is necessary to implement the proposed algorithms in autonomous machine-tending applications in a real-world industrial environment. Since the performance of deep learning models largely relies on the size of training data, a bigger and better dataset will be collected in the future and used to improve the proposed approach. Image deblurring [44–46] and reflection removal [47–49]
methods will also be applied to improve the quality of the captured images. Moreover, a reference dictionary containing common abnormal conditions of CNC machines will be designed to match and correct text recognition results.

6. Conclusions

Smart manufacturing has been considered high intelligence, efficiency, accuracy, productivity, and safety compared to traditional manufacturing. With robotics development, vision sensors and computer vision-based algorithms have been widely used in smart manufacturing to tackle complex and hazardous production tasks. Automatic working status recognition and emergency handling through autonomous robot systems have become critical steps in the autonomous manufacturing process, ensuring that machines work properly without human involvement. However, onsite workers still carry out the current inspection of the manufacturing process and machine tending by working with cobots because of lacking an appropriate intelligent manufacturing approach for machine detection and working status recognition. This paper developed an automatic deep learning-based approach to detect a particular CNC machine and the human-machine interface simultaneously in real-time in a complicated lab environment through a camera mounted on the mobile manipulator. In addition, it can identify the working status of the machine by automatically recognizing the text information on the HMI display. According to the validation results, the developed methods are proven to achieve good performances in both accuracy and efficiency. The proposed target CNC machine detection method is 16.5% more accurate than the milestone method Faster-RCNN. The developed machine’s working status recognition method is 10% more accurate than the benchmark algorithm tested. However, the performance of deep learning-based methods often largely relies on the size of training data and image qualities, so a bigger and better dataset will be collected in the future and used to improve the proposed approach.

Author Contributions: Conceptualization, F.J. and R.A.; methodology, F.J.; software, F.J.; validation, F.J.; formal analysis, F.J.; resources, F.J.; data curation, F.J.; writing—original draft preparation, F.J.; writing—review and editing, A.J., Y.M., and R.A.; visualization, F.J.; supervision, Y.M. and R.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NSERC (Grant No. NSERC ALLRP561048-20, and CRDPJ 537378-18) and the Minister of Economic Development, Trade and Tourism (through Major Innovation Project) for funding this project. The APC was funded by the NSERC (Grant No. NSERC ALLRP561048-20, and CRDPJ 537378-18) and the Minister of Economic Development, Trade and Tourism (through Major Innovation Project) for funding this project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors acknowledge the NSERC (Grant No. NSERC ALLRP561048-20, and CRDPJ 537378-18) and the Minister of Economic Development, Trade and Tourism (through Major Innovation Project) for funding this project.

Conflicts of Interest: The authors declare no conflict of interest.

References


