Hindawi
Mobile Information Systems
Volume 2021, Article ID 5555833, 10 pages
https://doi.org/10.1155/2021/5555833

Research Article

A Lightweight Application for Reading Digital Measurement and Inputting Condition Assessment in Manufacturing Industry

Jung-Sing Jwo,[1,2] Ching-Sheng Lin,[1] Cheng-Hsiung Lee,[1] and Chenhao Wang[3]

[1]*Master Program of Digital Innovation, Tunghai University, Taichung 40704, Taiwan*
[2]*Department of Computer Science, Tunghai University, Taichung 40704, Taiwan*
[3]*ZhiQi Railway Equipment Co. Ltd, Taiyuan 030032, China*

Correspondence should be addressed to Ching-Sheng Lin; cslin612@thu.edu.tw

Received 5 January 2021; Revised 28 February 2021; Accepted 19 March 2021; Published 28 March 2021

Academic Editor: Jong M. Choi

There is a vast need for the use of digital display instruments in the manufacturing industry due to the simple operation and high precision. In addition to the numerical data acquisition, it is usually necessary to input additional text for the condition assessment as well. However, since most of these measure instruments do not provide any interfaces for users to access the values and it often lacks proper devices to input the text during the working process, these two tasks are highly human intensive under current conditions. In order to facilitate the smooth running of the work for operators, we propose a lightweight application which can be installed on smartphones or wearable devices using multidigit recognition and speech recognition techniques without changing too much of their workflow. The experimental results demonstrate that our approach can achieve high accuracy. Thus, the proposed solution can effectively resolve data input issues in the manufacturing sites, thereby reducing human labor, increasing productivity, and automating part of the process. Taking operators' existing workflow into consideration for design can provide an application with an easy learning curve. Moreover, with the rapid and economical approach, companies can financially benefit from the development of this low-cost application, especially for small- and medium-sized enterprises.

1. Introduction

The Industry 4.0 concept [1], which is introduced to a wide audience initially during the Hannover Trade Fair in 2011, has promoted a vision of a new Industrial Revolution and encouraged the development of numerous disruptive technologies, such as cloud computing, Internet of Things (IoT), Big Data, and Artificial Intelligence (AI).

In most manufacturing sites, the use of digital measurement instruments is highly demanding since they are simple to operate and offer relatively high accuracy. However, these tools, in general, are not equipped with computer interfaces to acquire values for further processing. In addition, it is usually necessary to input additional text after inspecting the surface condition of the workpiece. During the working process, a proper device for operators to input the text is not easy to have. Although the growth of IoT has enabled the communication and exchange of information among devices, such profound changes to the enterprise system and environment often require large investments and can easily lead to shortfalls. According to the report [2], small- and medium-sized enterprises (SMEs) play an important role to represent over 99% of all enterprises in the European Union whereas the amount of large companies is less than one percent. Nevertheless, SMEs usually have lower digitalization level and gain limited access to the resource in comparison with the larger organizations. All these could pose the challenge to establish an interconnected industrial value creation to ensure future competitiveness and sustainability for SMEs. Thus, many manufacturing companies still rely on human labor for numerical data acquisition and text input. Therefore, providing a lightweight solution to address these two tasks for operators and taking their existing working conditions into consideration are crucial.

With the revolutionary advancement and the progressive development of neural network methods, AI has been

applied with great success in various fields and has its unique impact continuously. Due to its capability of solving complex problems, many companies in the industry have taken AI techniques to overcome the hurdles of their current methods and put them into the perspective of Industry 4.0 paradigm [3]. Visual and verbal communications are two of the most efficient and intuitive ways for human beings to work and learn. We will adopt two AI solutions, multidigit recognition and speech recognition, to resolve the issues of reading digital measurement and inputting condition assessment tasks. Since human technology is made by humans, for humans, it is important to take into account the human factor for digitalization challenges. In our application design, we take a step forward to include human-in-the-loop [4] design aspects in order to align the existing process and improve human's performance.

The contribution of this paper is threefold:

(1) Addressing the digitalization challenge in the manufacturing site: we propose a solution based on vision and speech recognition to extract the shop floor data for tackling the digitalization challenge in the manufacturing site. The experimental results of the proposed approach can exhibit a strong potential to be employed for automated data gathering.

(2) Developing a lightweight and low-cost application: the main goal of this study is to provide a feasible approach which could be used for the manufacturing industry, especially for SMEs. We implement our solution as a lightweight and low-cost app which is affordable for SMEs and could be easily extended to other smart devices such as wearable displays.

(3) Producing a human-in-the-loop design concept: given that human is the most important asset in the manufacturing industry, our design concept takes operators' existing workflow into consideration to provide an interface with an easy learning curve. The adjusted work procedure with the aid of smart technology could minimize the learning effort and increase productivity.

The remainder of this paper is organized as follows. Section 2 reviews state-of-the-art techniques related to the work of this paper. The proposed application is described in Section 3 which includes design concept and detailed program implementation. We explain the experimental setting and report results in Section 4. In Section 5, we present conclusions and discuss future research directions.

2. Related Work

In this section, we provide an overview of technologies related to this research including speech recognition, optical character recognition, and user experience.

Automatic Speech Recognition (ASR), which translates user's voice into text, aims at natural communication between humans and machines through languages. It is commonly supported in many applications of our daily life. There are a number of commercial and open-source packages which have been developed such as Microsoft API Speech, Google Speech API, and Sphinx-4 [5]. These three systems are evaluated on data selected from the TIMIT corpus [6] and ITU (International Telecommunication Union) using the word error rate (WER) metric [7]. The experiments report that Google API achieved 9% WER, Microsoft API achieved 18% WER, and Sphinx-4 achieved 37% WER.

In the early stage, hand-crafted features are widely designed for the application of optical character recognition (OCR), such as the exploitation of maximally stable extremal regions (MSERs) [8] for text detection and the combination of HOG with the Bag of Strokelets [9] for text recognition. Most recent approaches have shifted toward deep learning methods with a special interest in models which can be trained in an end-to-end manner [10–12]. Due to the increasing popularity of smartphones, several OCR applications have been developed for mobile devices with different purposes. Camera Reading for Blind People project [13] integrates OCR and text-to-speech synthesis (TTS) modules to build an iOS app in order to help blind users "read" text documents. Being unable to automatically align images restricts the usage of the app. Spot + OCR [14] is an Android app which communicates with users by a short vibration when the camera captures the text in the environment. The vibration reminds users to stabilize the phone, and subsequently, the Spot + OCR is able to run OCR program.

According to ISO 9241-210 [15], user experience (UX) is defined as "person's perceptions and responses resulting from the use and/or anticipated use of a product, system, or service." It is, especially, important in the manufacturing to consider the human-in-the-loop challenge to provide a good user experience for the purpose of achieving the task's goal successfully and efficiently. A rule of thumb for designing interactions is to ensure user participation, provide a natural and understandable collaboration, and avoid disturbing users [16]. The basis of the graphical user interface (GUI) is the fact that recognition is easier than recall [17]. Smartphone apps should offer services to first time users without requiring lengthy instructions and a steep learning curve [18].

3. The Proposed Lightweight Application

As the purpose of this research is to help operators reduce the workload and ease the task without changing their working procedures too much, we discuss the user experience-based design concept, followed by the implementation of mobile device applications and kernel recognition techniques. The case study is about the inspection of railway wheel and axle including dimension measurement of parts and visual examination of surface condition.

3.1. Design Concept. Since human technology is made by humans, for humans [4], it would be important to make sure that the proposed solution takes into account the operators' behavior and comfort level instead of only asking them to learn a new working procedure. We first observe the operators' workflow, conduct face-to-face interviews, and

correlate their verbal explanation with their actions performed in the working environment. The steps of their examination procedure are dissected as follows:

(1) Before inspecting a workpiece and performing the required activities, the operators have to print out the work orders and take with them. This is a very important step to know the identification and corresponding information of the workpiece beforehand.

(2) During the inspection procedure, the operators use either digital instruments to measure objects or human eyes to inspect the surface condition. Afterward, they need to write down what they read and observe by pen.

(3) After finishing the job, the operators have to go back to the office and input all obtained values and descriptions into the system. Although this step is crucial for digitization and data integration, they often tend to input altogether until all workpieces are examined in order to save more time. This would cause the data availability issue and duplicated work as those values and information have already been manually written down before.

By analyzing the above workflow, we propose a lightweight solution to assist operators in two perspectives, increasing data acquisition speed and reducing labor costs. We use smart technologies for operators to input data by speech recognition and digit recognition and adopt AI technologies to ease their tasks and improve their performance. Figure 1 shows the system architecture. By looking at the inspection item shown on the phone screen, the operators could use the default input approach for the given inspection item or select the preferred approach (voice or camera input). Once the result is obtained, the operators have to confirm the correctness and the data will be directly inserted into the backend system. The procedure will be repeated for the next inspection item. There are three advantages in this design. First, as we consider their current workflow, the operators can still maintain their familiar working routine without sacrificing too much of their existing strengths. Second, unlike the old scheme that the operators have to input all values while coming back to the desk, they can stay at the manufacturing site and be more productive. Finally, they only have to carry a mobile device with them rather than using paper and pen with a high risk of loss and damage.

3.2. Mobile Device App Implementation. The mobile device app serves as the means to link the user's input to the backend system. Tapping an icon on a touchscreen usually requires more attention to reach sufficient precision, and especially, operators have to wear gloves to avoid stains while performing measurements and examinations. It would therefore be a hurdle for them to touch screens during their working activities. Developing an intuitive and convenient experience in the manufacturing environment is necessary and urgent. To resolve the above restriction, we propose voice and camera inputs as two primary input approaches

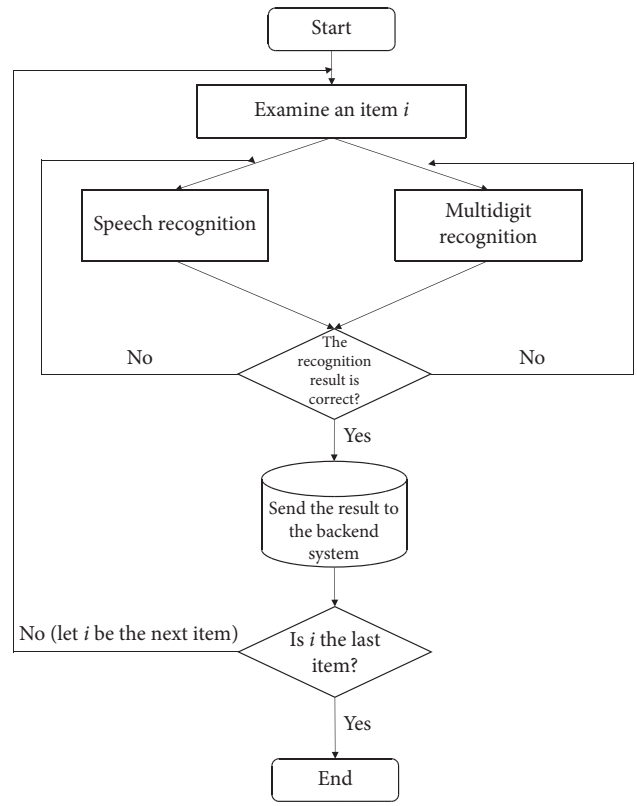
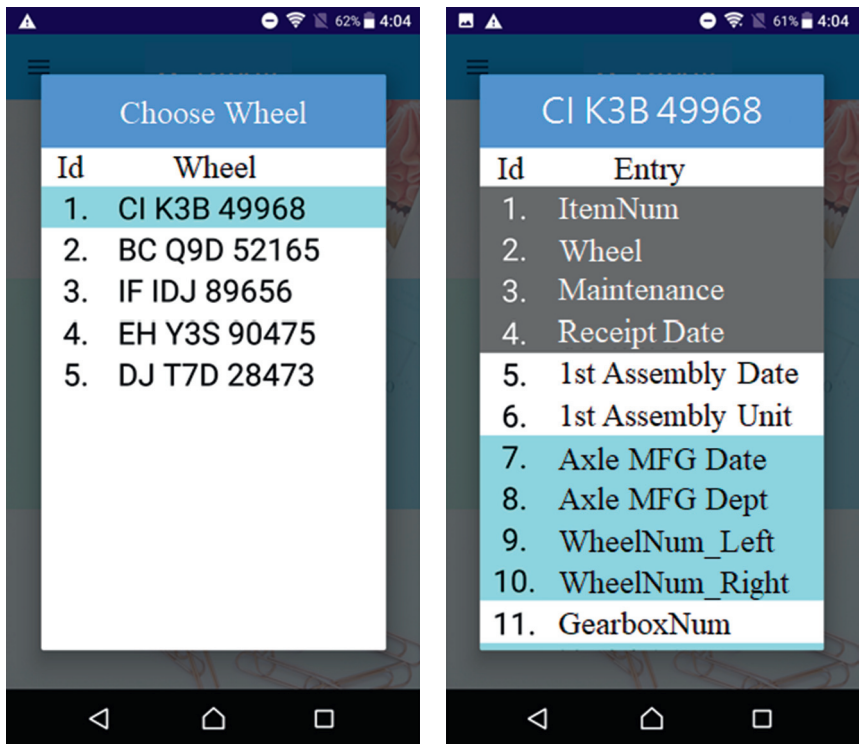


Figure 1: The system architecture of our approach.

but still keep the touchscreen option open. Moreover, to align with their existing routine, the page-to-page flow of our app follows their current working schedule so as to prevent their extra burden to learn new procedures. Due to the working conditions, we keep the interface design of the app simple to reduce distraction and enhance usability. The key functionalities of our proposed app along with the two major input methods, voice and camera approaches, are discussed as follows.

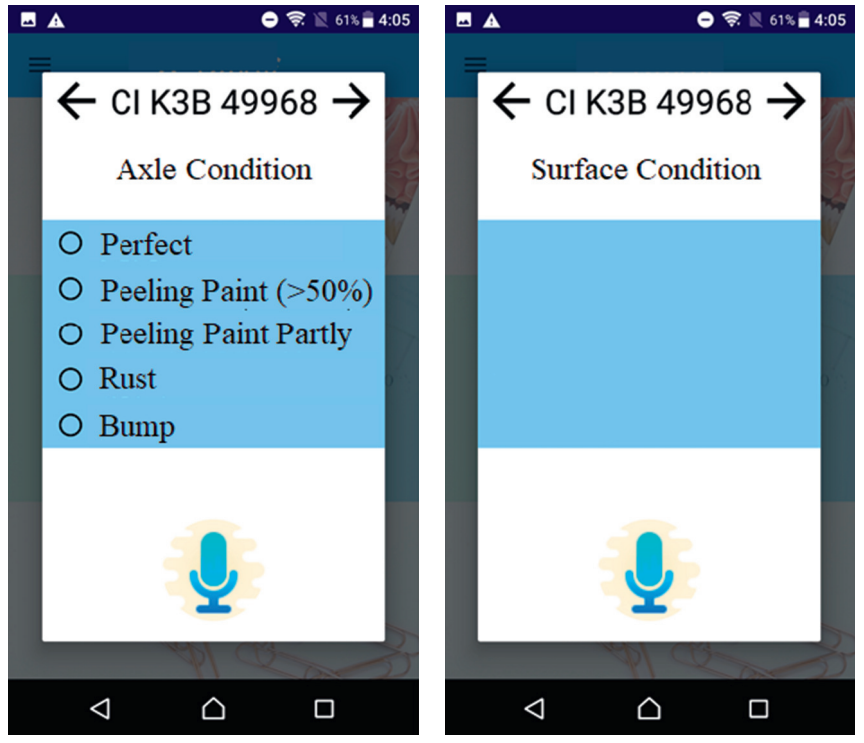
The main menu of the app is used to select from the available workpieces and is shown in Figure 2(a). Since the app is connected to the backend system, the data are always up to date. Operators can either use voice input or touch operation to make the choice. Once the operator has selected, the succeeding page in Figure 2(b) displays the information related to the prior selected workpiece which is a list restored from the previous checkpoint or a new checklist if the workpiece is not inspected before. Subsequently, the inspection page begins. There are two different kinds of inspection methods. One is to use digital display instruments to measure the dimensions of the railway wheel and axle in various positions, while the other is to evaluate the condition or damage on the wheel and axle surface by eye. We offer two AI-based approaches, speech and digit recognition, to automatically identify their measuring and observational results.

For the surface inspection which produces observational results, we provide voice input to interact with operators. There are two different modes in the app page: one is to say from predefined options (Figure 3(a)) while the other allows operators to utter free-text about additional information



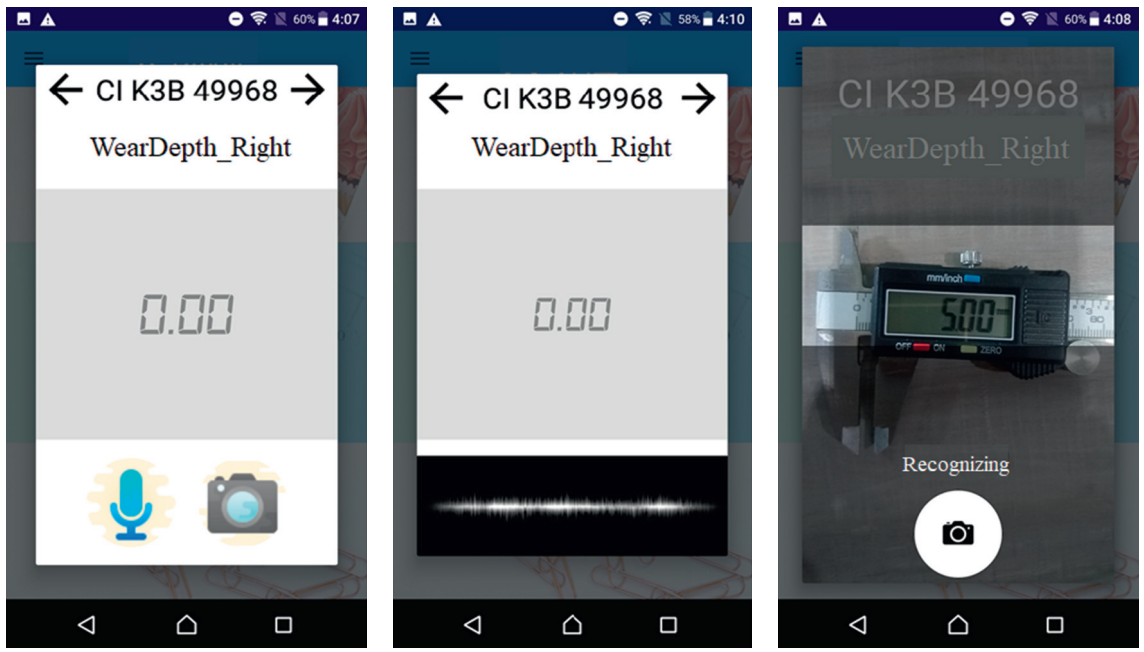
(a) (b)

Figure 2: (a) Workpieces page; (b) information page.



(a) (b)

Figure 3: Voice input for predefined text (a) and free-text (b).



(a) (b) (c)

Figure 4: Dimension measurement: (a) selection page, (b) voice mode, and (c) camera mode.

(Figure 3(b)). With the aid of voice input, operators are able to report their examination instantly and conveniently. On the other hand, regarding the dimension measurement (Figure 4(a)) which generates measuring results, both camera and voice inputs are provided for operators to use. In the voice mode (Figure 4(b)), it is similar to the surface inspection where operators say the number displayed on the digital measurement instrument. In the camera mode, operators only have to move the lens closer to the object and our app will automatically capture instantaneous images sequentially to carry out the digit recognition (Figure 4(c)). If a number is found, our app will return the recognition result. Otherwise, the image will be discarded and the image capturing process will be continuously running. Camera and voice control application enable users to input results without pressing buttons, making them ideal for use in the manufacturing site. We will discuss the speech and digit recognition techniques that we adopt in Section 3.3.

Once the recognition process is accomplished and the output is confirmed by the operator, the value or text will automatically be stored in the backend system. This is a very important functionality for the whole process because our app not only inserts data into the system but also retrieves the data from the system to display in the app page. There should be an intermediate interface to connect the app and the system. To achieve this, we implement a RESTful API as an intermediate bridge for the app to issue an HTTP GET or PUT request to access the existing backend system.

3.3. Recognition Methodology. There are two recognition systems used in our proposed app. The first one, speech recognition, is to transcribe spoken utterances to a sequence of words. In this paper, the context of utterance includes decimal number shown on the digital measurement instrument, predefined text described the object condition, and free-text uttered by operators for special notice. The second one, multidigit recognition, is to extract the number displayed on the digital display panel.

In the speech recognition implementation, we directly adopt two out-of-the-box applications (Google Cloud Speech API [19] and iFlytek [20]) with some modifications and compare the performance for our use. Google Cloud Speech API enables to convert speakers' voice to text through modern deep learning techniques and provides an easy integration for applications to recognize over 80 languages [21]. It has been reported to achieve high speech recognition accuracy for different tasks [22]. iFlytek, founded in 1999, is a Chinese technology enterprise specializing in speech intelligence. It provides Automatic Speech Recognition (ASR) API for various systems as well as supports for different programming languages. Meanwhile, many commercial apps, such as WeChat and Weibo, have used iFlytek as the voice input method [23]. In Mandarin Chinese, there are about 1300 syllables and more than 13000 characters. Since each character is pronounced as a syllable, approximately the average number of characters per syllable is around 10. It will suffer from homophone ambiguities and cause speech recognition errors. Therefore, instead of directly taking the recognition result from the service, we conduct postprocessing to reduce the error rate. The syllables of ten Chinese digits (from "0" to "9") are ("ling2," "yi1," "er4," "san1," "si4," "wu3," "liu4," "qi1," "ba1," and "jiu3"). If any speech character recognition result shares a syllable with one of the digits, we will convert the recognition result to the corresponding Chinese digit.

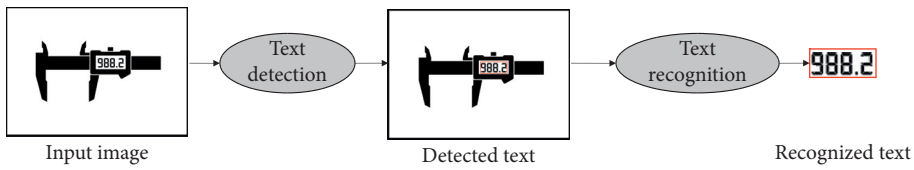


Figure 5: The flow of multidigit recognition.

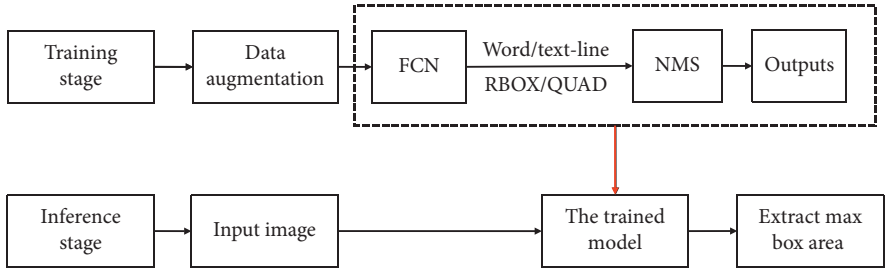


Figure 6: The training and inference flows of text detection.

In the multidigit recognition for optical measuring instruments, we propose a two-step method which consists of text detection and recognition. An overview of the pipeline of our approach is illustrated in Figure 5. Given an image which contains an optical measuring instrument, we first detect the text region which contains the number. Text detection is very critical to the pipeline as it is the premise for later text recognition. Then, the detected block is extracted and supplied to the text recognition module to determine the number.

East [24] is used as a text detector in our text detection module to extract the candidate number region. East detector consists of two important components, fully convolutional network (FCN) and nonmaximum suppression (NMS), and is depicted by the dashed rectangle in Figure 6. The FCN takes as input an image and generates word or text-line level predictions which could be either rotated box (RBOX) or quadrangle (QUAD). The above predicted text predictions are then sent to nonmaximum suppression (NMS) to produce the final output. To enhance the model generalization capability and overcome data scarcity in our application, some schemes, such as random cropping of nonnumber region and varying brightness and contrast, are used for data augmentation during the training phase. In the inference stage, we only select the output box with the maximum area as the result since we assume that the number should be the max text region in the image.

In text recognition, we combine the rectification network (RN) and sequence recognition network (SRN) to train a text recognizer [25]. The RN is used to rectify the distorted images and the SRN is to output the predicted text. The RN trains a weak learner to predict the offset of each part of the image and then applies sampling to rectify the image to be free of geometric constraints. Once the rectified image is generated, the SRN which is a CNN-BLSTM model followed by an attention-based decoder is to accurately learn the alignment between target label and characters in images [25, 26]. Regarding the training data volume, we also perform data augmentation to reduce human efforts and costs. For each experimental instrument, we manually crop five templates for each digit from zero to nine. As we have six instruments to experiment, there are 30 templates for each digit set and the "digit set of 0" is shown in Figure 7. To generate a number for training, we then are able to sample each digit of the number from the corresponding digit set and concatenate those sampled images as the generated image of the number. Figure 8 is an example of a generated number 823945. Note that since the decimal point of each digital instrument is fixed, we will omit the decimal point during the training and inference phases and simply put the decimal point back by the app program.

4. Experiment and Results

The performance evaluation of our proposed approach is conducted on two tasks: speech recognition and multidigit recognition.

4.1. Speech Recognition. There are two different datasets used for speech recognition, namely, the text and the number. In the text set, there are thirty utterances which contain predefined text of inspection checklist and free-text related to the surface conditions. In the number set, there are fifty decimal numbers. We invite five persons to participate in this experiment and compare the speech recognition accuracy on Google Cloud Speech API and iFlytek.

The detailed results are shown in Table 1. The overall accuracy rate of iFlytek is 91.6% which is higher than that of Google (83.3%). In general, these two systems all have better results in number than in text recognition. From the comparison in the text recognition, iFlytek achieves 86.6% accuracy which is slightly higher than Google (80.0%). About the comparison in the number recognition, iFlytek obtains 94.6% accuracy which is about 0.093 higher than Google. Note that iFlytek also supports to recognize 24 Chinese dialects.

Figure 7: Thirty templates of the digit set of 0.

4.2. Multidigit Recognition. For the multidigit recognition, we evaluate for both text detection and text recognition. In the following, we will discuss the tools, dataset, and metrics to be used for the assessment, followed by the experimental results and analysis.

There are six digital instruments used for our experiment as shown in Figure 9(a). Tool-1 is for the measurement of the distance between the backs of the wheel flanges. Tool-2 is for the measurement of the wheel set position. Tool-3 is for the measurement of the diameter of the wheels. Tool-4 is for the measurement of the journal diameter. Tool-5 is for the measurement of the diameter of the dust-proof set. Tool-6 is for the measurement of the wheel wear.

During the training phase, 50 images are taken from each instrument and augmentation techniques mentioned in Section 3.3 are used to generate additional 3600 images, resulting in a total of 3900 images to train the text detector. To produce training images for text recognizer, we apply the augmentation techniques discussed in Section 3.3 to create 20000 images. The evaluation metrics for text detection are recall, precision, and *F*-score. Recall measures how well the proposed approach is for retrieving correct regions, while precision indicates how accurate the technique is for predicting correct regions. In our experiment, the detected region will be considered correct if its intersection-over-union (IoU) with the ground truth is higher than 0.65. The *F*-score is defined to be the harmonic mean of recall and precision. Regarding the evaluation metric for text recognition, we use accuracy to measure the ability of the recognizer.

Figure 8: An example of a generated number.

Table 1: The accuracy comparison of speech recognition between Google and iFlytek.

Model	Google		iFlytek	
Type	Text	Number	Text	Number
Accuracy	0.800	0.853	0.866	0.946
Average	0.833		0.916	

In the testing stage, 100 images for each tool are used to evaluate the text detector, resulting in 600 images denoted as Testing_Set. We compare two models where one is pretrained on ICDAR [27, 28] (denoted as TD_Pretrained model) and the other is afterward fine-tuned using our augmentation data (denoted as TD_Fine-tuned model). For notational simplicity, we express those images extracted by TD_Fine-tuned model as Sys_Ext. As shown in Table 2 with IoU = 0.65, the overall *F*-score of the TD_Fine-tuned model is 100% whereas the TD_Pretrained is about 57%. In general, the TD_Fine-tuned model almost performs perfectly in the detection task. In addition to the IoU set to 0.65, we also use two thresholds, 0.75 and 0.85, to compare the results shown in Table 3. The performance is almost perfect under IoU = 0.75 while the average *F*-score drops to 0.73 under IoU = 0.85. A few of the detection results by TD_Fine-tuned model are illustrated in Figure 9(b).

The text recognition model is pretrained on two synthetic image datasets [29, 30] (denoted as TR_Pretrained model) and then fine-tuned using our augmentation data (denoted as TR_Fine-tuned model). There are two testing datasets, Sys_Ext and Hu_Ext, used to evaluate text recognition performance where Sys_Ext is the dataset mentioned in the last paragraph and Hu_Ext is the dataset that we manually extract the number section in the images from Testing_Set. The main purpose of the evaluation is to compare the performance between TR_Pretrained model and TR_Fine-tuned model and also investigate whether there is any significant performance difference between using Sys_Ext and Hu_Ext. The experimental results are presented in Table 4. It can be observed that the TR_Fine-tuned model significantly improves the average accuracy over the TR_Pretrained model on the Hu_Ext dataset which has been increased from 0.28 to 0.95. When comparing the results on the Sys_Ext dataset, TR_Fine-tuned model (0.95) also greatly outperforms TR_Pretrained model (0.24). For TR_Fine-tuned model, all accuracy values are above 0.9 and several values achieve 0.99. Overall, TR_Fine-tuned model can effectively utilize our augmented images to solve the text recognition problem and result in

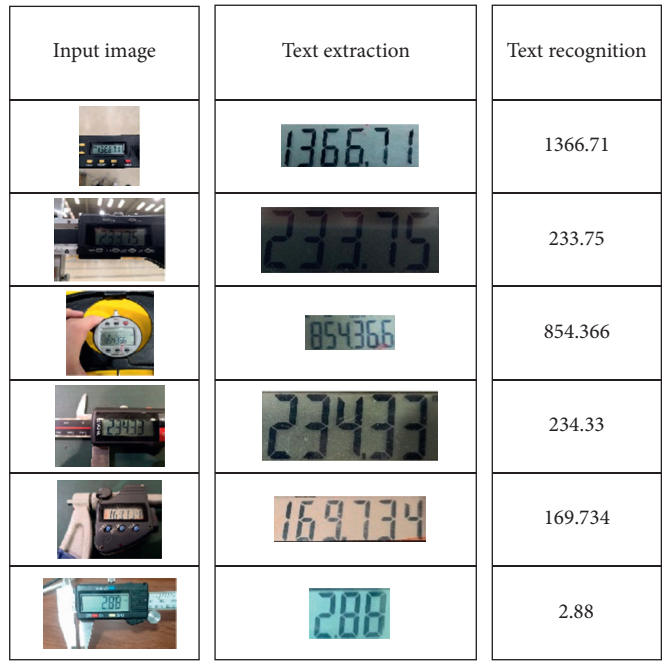


Figure 9: (a) The input images from Tool-1 to Tool-6. (b) The outputs of text detection. (c) The outputs of text recognition.

Table 2: The performance comparison of pretrained and fine-tuned models for six digital measurement tools in text detection.

	TD_Pretrained model			TD_Fine-tuned model		
	R	*P*	*F*	*R*	*P*	*F*
Tool-1	0.64	0.64	0.64	1	1	1
Tool-2	0.09	0.09	0.09	1	1	1
Tool-3	0.84	0.84	0.84	1	1	1
Tool-4	0.73	0.73	0.73	1	1	1
Tool-5	0.61	0.66	0.64	1	1	1
Tool-6	0.51	0.52	0.51	1	1	1
Average	0.57	0.58	0.57	1	1	1

Table 3: The detection comparison of *F*-score for different IoUs using the fine-tuned model in text detection.

	IoU = 0.65	IoU = 0.75	IoU = 0.85
Tool-1	1	1	0.75
Tool-2	1	0.99	0.84
Tool-3	1	1	0.75
Tool-4	1	1	0.47
Tool-5	1	1	0.74
Tool-6	1	1	0.83
Average	1	0.99	0.73

Table 4: The recognition accuracy comparison of pretrained and fine-tuned models for six digital measurement tools as well as the comparison with Aster and Clova.

	TR_Pretrained		TR_Fine-tuned		Aster_Pretrained		Aster_Fine-tuned		Clova_Pretrained		Clova_Fine-tuned	
	Sys_Ext	Hu_Ext	Sys_Ext	Hu_Ext	Sys_Ext	Hu_Ext	Sys_Ext	Hu_Ext	Sys_Ext	Hu_Ext	Sys_Ext	Hu_Ext
Tool-1	0.50	0.62	0.91	0.92	0.4	0.4	0.94	0.96	0.29	0.28	0.85	0.91
Tool-2	0.31	0.21	0.94	0.91	0.33	0.17	0.86	0.75	0.17	0.15	0.91	0.91
Tool-3	0.22	0.33	0.95	0.99	0.55	0.72	0.96	0.95	0.25	0.29	0.9	0.94
Tool-4	0.12	0.16	0.99	0.99	0.09	0.07	0.94	0.92	0.07	0.06	0.94	0.96
Tool-5	0.05	0.06	0.93	0.91	0.01	0.01	0.89	0.87	0.00	0.00	0.85	0.89
Tool-6	0.26	0.34	0.99	0.98	0.35	0.42	0.94	0.92	0.21	0.24	0.99	0.99
Average	0.24	0.28	0.95	0.95	0.28	0.30	0.92	0.90	0.17	0.17	0.91	0.93

high accuracy for multidigit recognition. Regarding the performance comparison between two datasets Sys_Ext and Hu_Ext, TR_Fine-tuned model reaches similar results (0.95 : 0.95). There is also no comparable difference while using TR_Pretrained model (0.24 : 0.28). We can conjecture that these two datasets are similar (i.e., the images extracted by our TD_Fine-tuned model are similar to those images in Hu_Ext extracted by humans). The comparison with two other techniques, Aster [26] and Clova [31], are also displayed in Table 4. It is observed that all approaches including TR_Pretrained model are not able to properly deal with the text recognition task but will obtain much better results after fine-tuning with our augmented data. Note that all tools (from Tool-1 to Tool-6) in our experiment use seven-segment displays. Several examples of recognition results are shown in Figure 9(c).

The experiments are conducted on the Windows system with an Intel(R) Core(TM) i7-8750H CPU, 8 GB RAM, and NVIDIA GeForce GTX 1050 Ti GPU. We obtain an average text detection time of around 0.05 seconds and the average text recognition time is about 0.09 seconds.

5. Conclusions

We develop an APP based on vision and speech recognition to extract the shop floor data for addressing the digitalization challenge of SMEs. Our proposed approach is executed on mobile devices which are affordable for SMEs and could be easily extended to other portable smart devices. We produce a design concept by taking operators' existing workflow into consideration and providing an intuitive interface with an easy learning curve.

Although the proposed application is applicable in the current situation, there are several challenges to be further addressed for future research to maximize the benefits of the proposed approach:

(1) Investigation of interaction tools: the current version of the interaction requires users to hold the smartphone to control. Although it is feasible for the current scenario, we would like to investigate more convenient tools to replace handheld interaction and expand the industrial value. Instead of handheld selection, we plan to work with operators to personalize smart wearable products, such as wristband, in order to achieve maximum user comfort and in accordance with their user experience [32].

(2) Use of in-house development: for our current speech recognition module, we use cloud-based service (i.e., iFlytek) which is applicable to several industries. However, based on our experiences of university-industry cooperation, many corporations require a high level of information sensitivity and avoid using cloud services. In further research work, our goal is to develop this component locally instead of uploading it to the cloud for recognition.

Data Availability

The data used to support this study have not been made available yet, as the supplier prevents this.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Pfeiffer, "The vision of "Industrie 4.0" in the making—a case of future told, tamed, and traded," *Nanoethics*, vol. 11, no. 1, pp. 107–121, 2017.

[2] J. Müller, L. Maier, J. Veile, and K. I. Voigt, "Cooperation strategies among SMEs for implementing industry 4.0," in *Proceedings of the Hamburg International Conference of Logistics (HICL)*, vol. 23, pp. 301–318, Hamburg, Germany, October 2017.

[3] J. Lee, H. Davari, J. Singh, and V. Pandhare, "Industrial artificial intelligence for industry 4.0-based manufacturing systems," *Manufacturing Letters*, vol. 18, pp. 20–23, 2018.

[4] D. Nunes, J. S. Silva, and F. Boavida, *A Practical Introduction to Human-In-the-Loop Cyber-Physical Systems*, John Wiley & Sons, Hoboken, NJ, USA, 2018.

[5] P. Lamere, P. Kwok, W. Walker et al., "Design of the CMU sphinx-4 decoder," in *Proceedings of the Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 2003.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*, National Institute of Standards and Technology (NIST), vol. 107, p. 16, Gaithersburgh, MD, USA, 1988.

[7] V. Këpuska and G. Bohouta, "Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx)," *International Journal of Engineering Research and Applications*, vol. 7, no. 3, pp. 20–24, 2017.

[8] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proceedings of the Asian Conference on Computer Vision*, pp. 770–783, Queenstown, New Zealand, November 2010.

[9] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: a learned multi-scale representation for scene text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4042–4049, Columbus, OH, USA, June 2014.

[10] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5238–5246, Venice, Italy, October 2017.

[11] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "Fots: fast oriented text spotting with a unified network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5676–5685, Salt Lake City, UT, USA, June 2018.

[12] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai, "Mask textspotter: an end-to-end trainable neural network for spotting text with arbitrary shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 532–548, 2019.

[13] R. Neto and N. Fonseca, "Camera reading for blind people," *Procedia Technology*, vol. 16, pp. 1200–1209, 2014.

[14] L. Neat, R. Peng, S. Qin, and R. Manduchi, "Scene text access: a comparison of mobile OCR modalities for blind users," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 197–207, Los Angeles, CA, USA, March 2019.

[15] ISO 9241-210: Ergonomics of human-system Interaction–Part 210: Human-Centred Design for Interactive Systems (2010).
[16] M. Gil, M. Albert, J. Fons, and V. Pelechano, "Engineering human-in-the-loop interactions in cyber-physical systems," *Information and Software Technology*, vol. 126, p. 106349, 2020.
[17] J. Johnson, *Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Guidelines*, Elsevier, Amsterdam, Netherlands, 2013.
[18] T. Page, "Skeuomorphism or flat design: future directions in mobile device User Interface (UI) design education," *International Journal of Mobile Learning and Organisation*, vol. 8, no. 2, pp. 130–142, 2014.
[19] Google cloud platform—google cloud speech API. [Online]. Available: https://cloud.google.com/speech/.
[20] iFlytek. [Online]. Available: http://www.xfyun.cn/services/voicedictation.
[21] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in parkinson's disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22199–22208, 2017.
[22] T. Baumann, C. Kennington, J. Hough, and D. Schlangen, "Recognising conversational speech: what an incremental ASR should do for a dialogue system and how to get there," in *Dialogues with Social Robots*, pp. 421–432, Springer, Singapore, 2017.
[23] X. Yuan, Y. Chen, Y. Zhao et al., "Commandersong: a systematic approach for practical adversarial voice recognition," in *Proceedings of the 27th {USENIX} Security Symposium ({USENIX} Security 18)*, pp. 49–64, Baltimore, MD, USA, August 2018.
[24] X. Zhou, C. Yao, H. Wen et al., "East: an efficient and accurate scene text detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5551–5560, Honolulu, HI, USA, July 2017.
[25] C. Luo, L. Jin, and Z. Sun, "Moran: a multi-object rectified attention network for scene text recognition," *Pattern Recognition*, vol. 90, pp. 109–118, 2019.
[26] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: an attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2018.
[27] D. Karatzas, F. Shafait, S. Uchida et al., "ICDAR 2013 robust reading competition," in *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*, pp. 1484–1493, Washington, DC, USA, August 2013.
[28] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou et al., "ICDAR 2015 competition on robust reading," in *Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1156–1160, IEEE, Tunis, Tunisia, August 2015.
[29] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," 2014, http://arxiv.org/abs/1406.2227.
[30] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2315–2324, Las Vegas, NV, USA, June 2016.
[31] J. Baek, G. Kim, J. Lee et al., "What is wrong with scene text recognition model comparisons? dataset and model analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4715–4723, Seoul, Korea, October 2019.
[32] P. Zheng, H. Wang, Z. Sang et al., "Smart manufacturing systems for Industry 4.0: conceptual framework, scenarios, and future perspectives," *Frontiers of Mechanical Engineering*, vol. 13, no. 2, pp. 137–150, 2018.