Technology Explanation

Development of Stamp Inspection Technology Using AI and Construction of MLOps Platform

MITSUO Takashi・SUZUKI Kento・MIYAUCHI Yuuki・KIKUCHI Takayoshi

Abstract

In recent years, the automation of visual inspection of product appearance has transitioned from a rule-based approach to a technique utilizing AI (Artificial Intelligence). Kayaba aims to utilize AI to develop inspection technology that is robust against individual differences in products and changes in the inspection environment.

Shock absorbers, our main product, are inspected visually for all stamps, so there is a demand for labor-saving inspection work and improved quality assurance by preventing inspection errors. Therefore, by applying AI, we aim to develop a highly robust stamp inspection technology, thereby reducing labor and improving quality assurance.

In inspections using AI, changes in the environment over time cause the AI model to deteriorate and inspection accuracy to decrease. Regular relearning is necessary to maintain the system, but doing it manually places a heavy burden on administrators. In addition, there have been problems with the consistency and reproducibility of the results. Therefore, we have newly built an operation management system aimed at reducing the burden on developers and improving the efficiency of data management and operation.

This paper describes the development of the stamp inspection technology and the newly constructed system.

1 Introduction

KYB identifies its shock absorber (SA) products with a stamped product number, which is subject to 100% visual inspection. To improve quality assurance and reduce labor, the visual inspection needs to be automated. A rule-based automation approach could hardly maintain consistent inspection performance due to variations in the brightness (contrast) of inspection images caused by individual product differences or changes in the inspection environment. To solve this problem, we have applied Artificial Intelligence (AI) to develop a highly robust[Note 1)] technology for identifying stamped product numbers.

In order to apply the AI-based inspection technology to our factories for efficient operation, we need to increase the efficiency in the AI execution cycle (data collection, AI training, accuracy evaluation, model deployment[Note 2)]) and properly manage and operate the data, thereby maintaining the AI quality at a high level. Therefore we constructed a platform to implement MLOps[Note 3)].

The following chapters describe the stamp inspection technology and the MLOps platform.

Note 1) Refers to the state of a product whose characteristics are not susceptible to changes caused by external factors.

Note 2) The process of integrating a machine learning model into an existing environment (equipment).

Note 3) An initiative to make AI easier. A word combination created from AI Machine Learning (ML) and DevOps, a continuous development approach in the software field.

2 Development of Stamp Inspection Technology

2. 1 Inspection Target

The inspection target is the product number letters stamped on the circumference of the SA cylinder (Photo 1). The inspection is performed in the final process after painting to check the stamped product code for number errors and letter fading.

Photo 1 Inspection target

2. 2 Configuration of Stamp Inspection Machine

Fig. 1 shows the configuration of the developed stamp inspection machine. The inspection target is rotated and continuously captured by a line scan camera (which continuously captures images by linear motion and processes them to obtain images as shown in Photo 2). The machine identifies the stamped product number from the captured image and determines the pass/fail of the stamp.

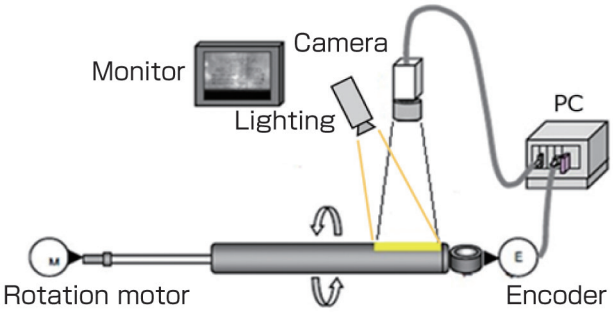

Fig. 1 Configuration of stamp inspection machine

Photo 2 Captured image

2. 3 Requirements

The following are the requirements for developing efficient, accurate stamp inspection technology:

[1] Establish a method for preparing efficient AI training images.

[2] Establish a character recognition algorithm that can maintain high recognition accuracy.

2. 4 Overview of Stamp Inspection Technology

2. 4. 1 Establishing a Method for Preparing Efficient AI Training Images

To build an AI model with high character recognition accuracy, the AI needs to learn images with various characteristics, including individual product differences and variable inspection environments. When collecting such training images in the initial development stage, two problems arose:

[1] It was difficult to uniformly collect training images of the 36 characters to be recognized, consisting of numbers, alphabets, and symbols, in a short period of time because these characters were not necessarily uniformly used in the actual products.

[2] It was difficult to collect in a short period of time training images of characters with different letter thicknesses caused by variable image brightness due to the state of surface painting or by the degree of wear of the stamping piece (a jig used to stamp the product number) (Photo 3).

To solve these problems, based on the CAD drawing (Fig. 2), we uniformly created hypothetical product number images (artificial images) for the 36 characters with different noise levels, including different letter thicknesses and color shading (Fig. 3). It was then possible to combine several tens of thousands of the artificial images with a small number of images of actual products to provide a set of training images, thereby creating a highly accurate AI model. Thus, we successfully established a method for collecting efficient training images to reduce the initial development time.

Photo 3 Images of stamped product number that look different

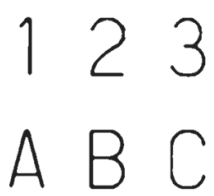

Fig. 2 Font data (sample)

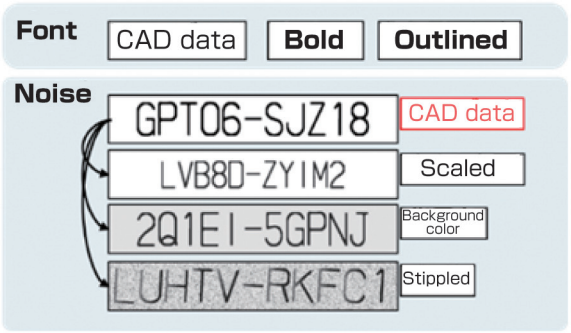

Fig. 3 Artificial image (sample)

2.4.2 Character Recognition Algorithm

In fact, the AI model, despite the introduction of the retraining approach[Note 4], produced stamped product number identification results that did not meet the target agreed with the relevant factory. Then, in addition to the retraining approach, the ensemble approach was introduced. In the ensemble approach, two or more AI models as shown in Fig. 4 are used to provide multiple results, from which a final result is output based on majority rule. In this project, the decision based on majority rule is made by YOLO[Note 5], CNN[Note 6], and an appearance check by a human operator (Fig. 5).

Specifically, YOLO and CNN recognize the stamped product number and determine if it matches the correct number provided. If both AI models answer the correct number, the product is accepted (pass). If both answer incorrect numbers, the product is rejected (fail). If one answers a wrong number (pending), the product is visually inspected by a human operator to determine pass/fail. Using this algorithm, we have established a method to identify anomalous products while keeping the false answer rate at a low level.

Note 4) An approach to re-training the AI by adding images of products that the AI failed to correctly recognize to the training dataset.

Note 5) An acronym for You Only Look Once. An algorithm used to identify "what's where in the picture".

Note 6) An acronym for Convolutional Neural Network. An algorithm used to identify "what's in the picture".

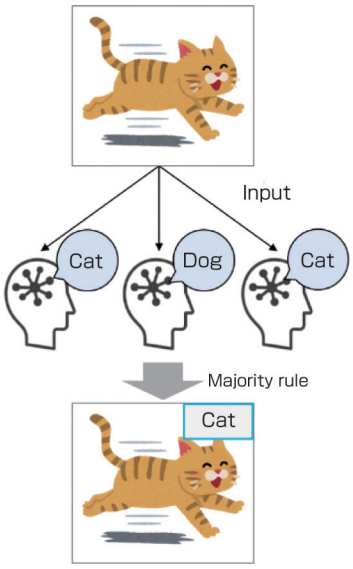

Fig. 4 Ensemble approach

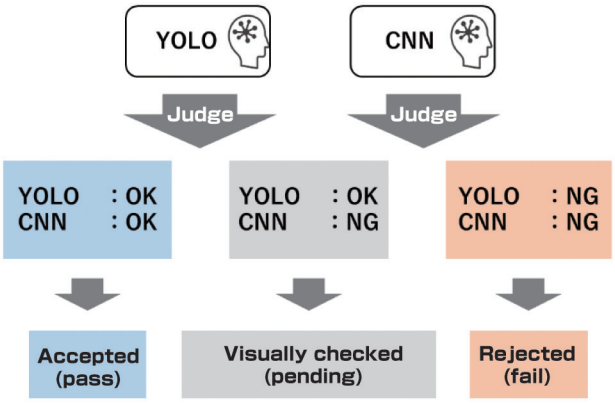

Fig. 5 Character recognition and evaluation flow

3 Construction of MLOps Platform

3. 1 Necessity of MLOps Platform

During the operation and management of the machine learning models technically built in the previous chapters, three major problems were identified. One was "complexity in setting up environments", such as a programming environment necessary for machine learning. Another was "difficulty in quality management", that is, the dependence of machine learning model management on users made it difficult to manage the quality of the machine learning models. The last one was "too many man-hours for operation and management", that is, the operation and management of the machine learning models required its administrators to spare many man-hours. To solve these problems, it was necessary to have an MLOps platform to efficiently operate the machine learning based system.

There are two major options for building an

MLOps platform: on-premises[Note 7] and cloud. Until now, the platform has been running on-premises. However, we faced several challenges. One example was "increased man-hours for maintaining model accuracy", which was necessary for compiling inspection results and re-evaluating AI models. Another example was the "ambiguous data and AI model management process" that did not ensure consistency and reproducibility of results. Faced with these challenges, we considered it advantageous to build a system in the cloud from the viewpoints of functional expandability, maintenance, future deployment, and operability. Then, we built a MLOps system series from data collection to model deployment in the AWS cloud. Fig. 6 shows an overview of the MLOps platform we built.

Note 7) A mode of operation in which the company owns and manages the software and hardware to build a system itself.

3. 2 Requirements

The following are the requirements for operating the MLOps platform:

[1] A flexible system that can also be used for applications other than the stamp inspection machine.

[2] Automate routine tasks that do not require human decision making to significantly reduce operational man-hours.

[3] Automate the human approval process to determine when to move to the next step.

3. 3 From Edge Inference to Data Storage

The platform needs to use different types of data, including retraining data (inference images that were misidentified by AI models) and inference logs. These types of data come in a variety of formats, including images and CSV. In addition, since it must be possible to use the platform for applications other than the stamp inspection machine, extensibility is required to be able in future to collect data in formats other than those currently compatible. Considering these situations, it was decided to use Amazon S3 [Note 8] to store data in a variety of formats.

It was also decided to transfer the data generated in the edge terminals[Note 9] to Amazon S3 via an intermediate server instead of uploading it directly. The reasons for this decision were twofold:

[1] Concerns about communication with external networks (instability).

[2] Concerns about the operational stability of the edge devices (no production impact).

The intermediate server must have an authentication to allow uploading to Amazon S3. It was decided to use AWS IoT Core[Note 10] for authentication management. The service allows only devices registered with AWS (in this case, the intermediate server) to use the temporary authentication to allow secure data upload.

Note 8) A storage service that can store and protect data regardless of type or capacity.

Note 9) Terminal devices connected to a network to collect and process data. For the purpose of product stamp inspection, they are responsible for a series of processes from camera capture to stamp identification with AI models.

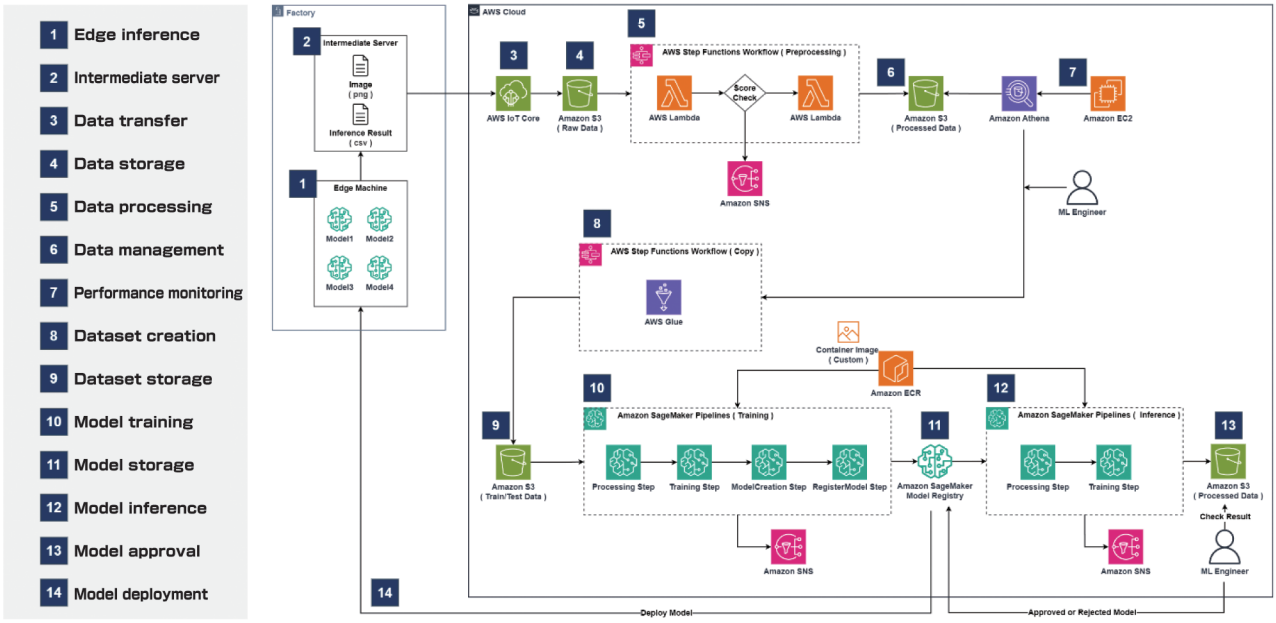

Fig. 6 Overview of MLOps platform

Note 10) A service that connects IoT devices to various AWS services.

3.4 From Data Processing to Performance Monitoring

Traditionally, human administrators would periodically determine if the AI models needed to be retrained. This way of working puts a heavy burden on the administrators and may not work in the future when the AI models need to cover more machines. Therefore, we established an automatic evaluation system to determine whether the AI models have degraded in performance. The evaluation is based on the following two criteria and determines whether the performance is below a threshold. Criterion [1] detects an abrupt decrease in the performance of the AI models. Criterion [2] detects a decline in AI model performance over time.

[1] Evaluate model performance on the day of processing (daily check).

[2] Evaluate model performance over the last five business days (weekly check).

The check results are notified to the Microsoft Teams chat via Amazon SNS[Note 11] (Fig. 7). The notification allows the administrators to identify the need for retraining, which significantly reduces the operational burden.

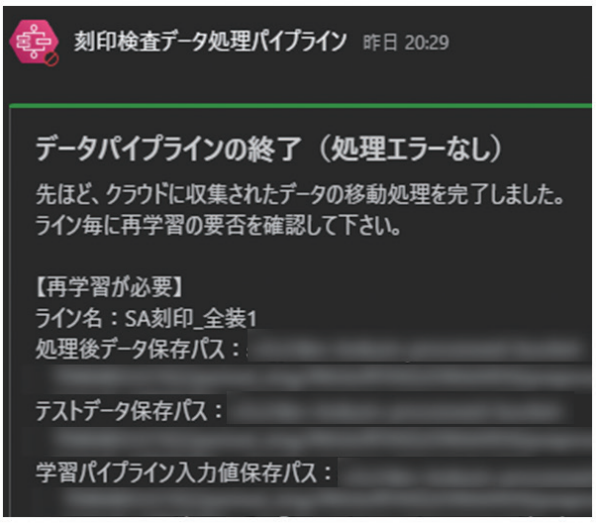

Fig. 7 Example of data processing completion notification

In addition, a dashboard was created using a BI tool[Note 12] to identify the routine performance and availability of the AI models (Fig. 8). The BI tool is Tableau[Note 13], which is provided as a data analysis environment across the company. On the dashboard, the administrators can not only view the performance and availability of the AI models, but also perform simplified ad hoc analysis[Note 14].

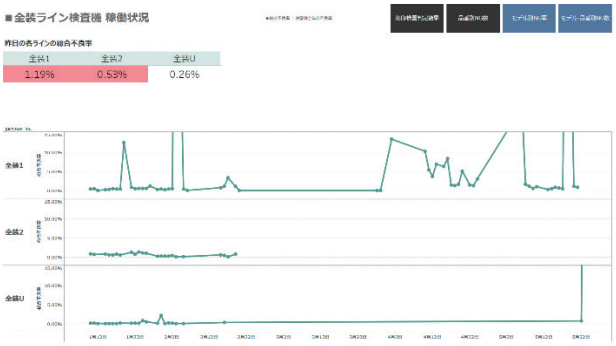

Fig. 8 AI model performance monitoring dashboard

Note 11) A service for exchanging messages between applications and different AWS services.

Note 12) A tool that compiles necessary information from large amounts of data accumulated in a company and analyzes it to be used effectively for management and operations.

Note 13) An application designed specifically for visualizing large amounts of complex data.

Note 14) A type of data analysis with no fixed analysis items or target that is performed as a one-time operation on an as-needed basis.

3.5 From Dataset Creation to Model Deployment

The processes from dataset creation to model deployment, which are the critical processes of the MLOps platform, are performed as follows:

[1] Create training/test datasets

[2] Retrain and approve AI models

[3] Deploy AI models

3.5.1 Creating Training/Test Datasets

When the function described in Section 3.4 deems it necessary to retrain the AI models, the creation of datasets is performed. To create the dateset, it was necessary to add data that had been manually annoutated[Note 15] to the source dataset. We automated the creation of the source dataset. The automated system notifies the Microsoft Teams chat when it is complete. Upon notification, the administrators can simply add the annotated data to complete the dataset creation process.

Note 15) A task of providing the correct response information needed to train an AI model.

3.5.2 Retraining and Approving AI Models

As an implementation environment for fully-managed[Note 16], high-speed learning, and inference, we use Amazon SageMaker[Note 17]. In particular, Amazon SageMaker Pipelines[Note 18], which is one of its functions to build ML pipelines[Note 19], is used to fulfil the purpose of this project, which is to reduce the operational burden of the administrators by building an MLOps platform. The pipelines can relatively easily automate the

series of MLOps processes, including model training, testing, registration, and deployment.

The stamp inspection system uses a total of four AI models, including the one for pre-processing. When one of the four models is retrained in the ML pipelines, it is necessary to check the inference results of the whole system. So, it was decided to have four kinds of performance evaluation in the inference pipelines as shown in Fig. 9. In this way, we have built a mechanism that can not only improve the performance of a single AI model but also evaluate the whole system.

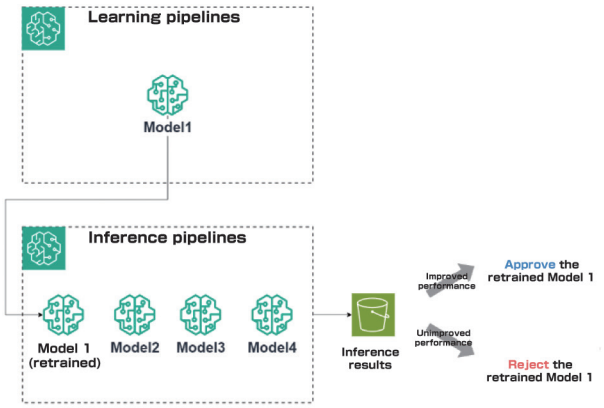

Fig. 9 Learning and inference pipeline processing flow

The system is designed so that when the processing reaches the output of the inference results, it notifies the Microsoft Teams chat of a request to check the results (Fig. 10). This notification allows users to determine the pass/fail of the AI model by simply interacting with the Microsoft Teams user interface (UI), such as "approve the retrained model" if the performance has improved, or "reject the retrained model" if the performance has not improved. The approved model is registered in the Amazon SageMaker Model Registry[Note 20]. It is then stored in Amazon S3 for deployment to edge devices.

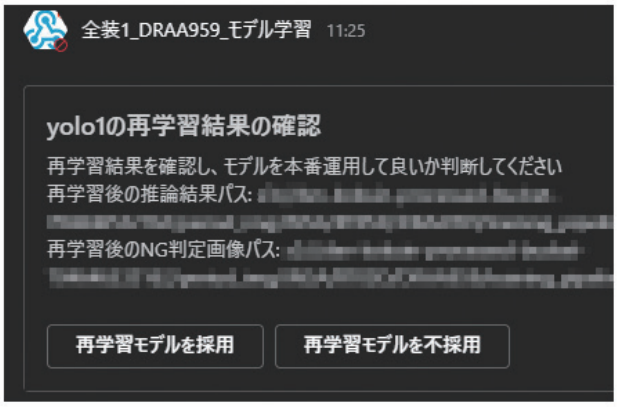

Fig. 10 Request to check inference results

Note 16) A service mode in which the contractor performs "operational tasks" such as "computer failure monitoring, data backup, and software version control" for the user.

Note 17) A service that provides an environment for rapidly developing, training, and deploying AI models.

Note 18) A workflow orchestration service that automates MLOps steps.

Note 19) A workflow of a series of steps/processes from data preprocessing, model training, evaluation, and deployment.

Note 20) A service that supports the management and deployment of AI models.

3. 5. 3 Deploying AI Models

For the same reason as the concerns about the data upload mentioned in Section 3.3, it was decided to deploy the AI models through an intermediate server. The intermediate server checks daily where the models are stored in Amazon S3. If a current model exists, it is temporarily downloaded to the intermediate server and then deployed to the edge terminals at a time that does not impact production.

3. 6 Retraining Process

To implement the retraining process described in Section 3.5, it was necessary for the administrators to perform data download from Amazon S3, and other tasks based on commands (CUI[Note 21]). In order to simplify such tasks, we created applications that can be operated on a GUI[Note 22] as shown in Fig. 11, which enables easy retraining. This has successfully reduced the manual work required for retraining.

Fig. 11 Retraining GUI

Combining these applications with the AWS cloud has improved the efficiency of the AI run cycle and ensured proper data operation and management, allowing AI quality to be maintained at a high level.

Note 21) A method that allows users to interact with (or operate) the computer by entering text commands.

Note 22) A method that allows users to interact with (or operate) the computer using visual elements such as icons and menus.

4 Future Prospects

The stamped product number identification technology has been introduced in four mass production lines at a single site as of June 1, 2024. As many of KYB's mass production lines still rely on 100% visual inspection by human operators to identify the stamped product number, we will continue to roll out this technology laterally to other sites, contributing to labor savings for operators and improved quality assurance.

Since we have established the AI model maintenance and management system, it is now possible to easily identify the reduction of man-hours for retraining AI models and the status of their operation. Repeated application of the system to mass production has achieved reliability, although the short operation time alone has not yet yielded a substantial result. For efficient deployment of the system, it is essential to cooperate with various departments in the factories. We will promote the introduction of the system by discussing the operation aspect of the system with the departments in charge.

We also plan to apply the AI-based inspection technology to processes other than the inspection of stamped product numbers. By promoting the deployment of the MLOps platform to such processes, we will contribute to higher productivity and product quality.

5 Concluding Remarks

This development project has achieved the labor saving of stamp inspection and the improvement of quality assurance.

Finally, we would like to take this opportunity to express our sincere gratitude to those from related departments who have provided substantial support and cooperation in this development project.

Authors

MITSUO Takashi

Joined the company in 2006.
R&D Sect. No.2, Production Technology R&D Center, Engineering Div.
Engaged in development of inspection and measurement technologies.

MIYAUCHI Yuuki

Joined the company in 2017.
Digital Strategy Office, Digital Transformation Improvement Div.
Engaged in building cloud-based systems.

SUZUKI Kento

Joined the company in 2016.
R&D Sect. No.2, Production Technology R&D Center, Engineering Div.
Engaged in development of inspection and measurement technologies.

KIKUCHI Takayoshi

Joined the company in 2014.
Digital Strategy Sect., Digital Transformation Improvement Div.
Engaged in building cloud-based systems.